THE ANNALS
of
MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

VOLUME XV

1944

THE ANNALS OF MATHEMATICAL STATISTICS

The Annals of Mathematical Statistics is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore 2, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the Annals of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore 2, Md., or to the Secretary of the Institute of Mathematical Statistics, P. S. Dwyer, 116 Rackham Hall, University of Michigan, Ann Arbor, Mich.

Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the Annals of Mathematical Statistics should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the Annals is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

Composed and Printed at the
WAVERLY PRESS, Inc.
Baltimore, Md., U. S. A.

THE ANNALS *of* MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE
OF MATHEMATICAL STATISTICS

Contents

Vol. XV, No. 1 — March, 1944

THIS JOURNAL IS
MANUFACTURED UNDER WARTIME CONDITIONS
IN CONFORMITY WITH ALL GOVERNMENT
REGULATIONS CONTROLLING THE USE OF
PAPER AND OTHER MATERIALS

THE ANNALS OF MATHEMATICAL STATISTICS

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore 2, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore 2, Md., or to the Secretary of the Institute of Mathematical Statistics, P. S. Dwyer, 116 Rackham Hall, University of Michigan, Ann Arbor, Mich.

Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879

g

11- 1

11-18-49

ON THE THEORY OF SYSTEMATIC SAMPLING, I

By William G. Madow and Lillian H. Madow[1,2]

1. Introduction. It is no longer necessary to demonstrate a need for the theory of designing samples. Many of the policy and operating decisions of both government and private industry are based on samples. There has been an increasing tendency in government and industry to make use of sampling theory.[3]

Unfortunately there are still considerable differences between the theory and practice of sampling. The origins of these differences are, on the one hand, the ignorance of administrators concerning the practical contributions that sampling theory can make, and on the other, the lack of sampling theory permitting the evaluation of some useful sampling designs.

Much has been and is being done towards bringing theory and practice into agreement.[4] Administrators and samplers are each successfully educating the others. However, there still exist sampling designs for which an adequate theory has not been developed, even though experience indicates that if such a theory were developed it would demonstrate the superiority of those designs over others for which a theory has been developed.

Perhaps the major omission of sampling theory today is the lack of any statistical method for reaching a decision on whether to take a completely random sample of n elements of a population of N elements, or to take a systematic sample, that is, to begin with element i, and select elements $i, i + k, \cdots, i + (n - 1)k$, as the sample, the starting point i being chosen at random and $N = kn$ approximately.[5] It is with respect to this question of whether to take a systematic[6] or random sample that the statistician is in a dilemma because he has the alternative of recommending a systematic sampling procedure for which no theory exists, or a random sampling procedure that may well yield worse

[1] Bureau of Agricultural Economics and Food Distribution Administration, U. S. Department of Agriculture, Washington, D. C.

[2] Presented at a meeting of the seminar in statistics of the Graduate School, U. S. Department of Agriculture, November 2, 1943.

[3] The recognition of the need for statisticians who know sampling theory has resulted in courses in sampling being given in some of the colleges and universities.

[4] One need only refer to the recent development of positions, the duties of which include giving advice on sampling techniques as well as working in the field of application.

[5] In this paper we will assume that $N = kn$. To do away with that assumption would not add much in the way of generality while it would require some fairly detailed discussion. It may be remarked that when N is not exactly kn, then systematic sampling procedures in which all starting points have equal probability of selection are biased, although the bias is usually trivial. If N is known this bias can be removed by sampling proportionate to possible size of systematic sample.

[6] As we define systematic sampling procedures, a systematic sampling procedure is a random sampling procedure in which many of the C_n^N selections of n from N items are excluded.

results than the systematic procedure. The purpose of this paper is to resolve that conflict by providing an adequate theory of systematic sampling.

In the following sections we present the first parts of our research in the theory of systematic samples. Although this research covers both the theory of sampling single elements and sampling clusters of elements, we shall consider, in this paper, the sampling units to be single elements, not clusters of elements. The latter problem will be dealt with in a later paper. We shall present the theory of systematic sampling both from an unstratified population and a stratified population. Formulas for the mean value and variances of the estimates are derived. Comparisons with random and stratified random sampling designs are made. Furthermore, the estimates of the variances and formulas for "optimum" size and allocation of samples are derived.

A fundamental part of the analysis is the demonstration that from a knowledge of the variance of the population[7] and certain serial correlations or serial variances, can be estimated the variance of estimates based on systematic samples. The basic results are:

a. if the serial correlations have a positive sum, systematic sampling is worse than random sampling,

b. if the serial correlations have a sum that is approximately zero, systematic sampling is approximately equivalent to random sampling, and

c. if the serial correlations have a negative sum, systematic sampling is better than random sampling.

2. The use of a finite population. In this paper we assume, for the calculation of the expected values, that we are sampling from a finite population of elements even though the size of the population may be large enough to permit the use of limiting distributions. Often, this is, mathematically, a matter of choice. The same results would be obtained by assuming a correctly defined multivariate normal distribution and using the notions of conditional probability. From a physical point of view, however, there are several factors that lead to the use of the finite population. We are most frequently sampling an existing population whose laws of transformation are either unknown or not mathematically expressed.[8] Consequently, the notion of a normal or other specified distribution from which we sample and use conditional probability is not part of our thinking concerning the physical problem. On the other hand, if we consider the population to be a finite population, and use a table of random numbers to draw our sample from the finite population, we are using only mathematics implicit in our physical problem. Furthermore, we do obtain a repeatable experiment, that of selecting a random number, that we know is in a state of statistical control.

In the usual problems of the theory of random sampling, the number of

[7] By "variance of population" without further qualification is meant the variance of a random sample of one element of the population.

[8] In other words, our population is not in a state of statistical control over time.

possible samples yielding different sample means is large enough so that the sample means may, with a sufficiently large size of population and sample, be expected to be approximately normally distributed. In systematic sampling, however, the number of possible sample means is usually very small and even if the sizes of population and sample are large, it is difficult to assume a normal distribution. Consequently, in our interpretation of the means and variances of systematic samples we are led to regard the elements of our populations as being the results of single observations on random variables, the distributions of which may vary from element to element. The interpretations that we then make become interpretations of conditional probability, and if the sizes of population and sample are sufficiently large, we can assume that the arithmetic mean of each of the possible sample means is normally distributed.

The theory of systematic sampling under the assumption of an appropriate normal multivariate distribution will be dealt with at a later time.

3. Definitions. Let the finite population to be sampled consist of N elements, $x_1, \cdots, x_N$.

By a sample design is meant the combination of a method of classifying these N elements into k classes that may or may not overlap, and a method of selecting one of these k classes, each class having a designated probability of being selected. The sampling procedure associated with a given sample design is the operation of selecting one of the k classes according to the method stated in the sample design. The sample is the particular class obtained by the sampling procedure.

By a random sampling procedure is meant any sampling procedure such that if the sampling design yields k classes then the probability of selecting anyone of these classes is $1/k$. Any sample design having a random sampling procedure associated with it is a random sampling design. One of the nonrandom sampling procedures that is being used is the procedure in which the classes have associated to them numbers, called sizes, and the probability of a given class being the sample is proportionate to its size.[9] Other nonrandom sampling procedures are doubtless being used.

By an unrestricted random sampling design for selecting n elements from N elements is meant the sampling design such that there are C_n^N classes, the possible selections of n from N elements, each having a probability of $1/C_n^N$ of being the sample. The associated random sampling procedure might consist in identifying each class by a number i, $i = 1, \cdots, C_n^N$ and selecting a number i from a table of random numbers. The random sampling procedure might also consist in identifying the N elements with numbers $j = 1, \cdots, N$, and then selecting a number j from a table of random numbers, then selecting a different number j from a table of random numbers, and following that procedure until n numbers

[9] For a discussion of this problem see the paper entitled, "On the theory of sampling from finite populations," by Morris H. Hansen and William N. Hurwitz, *Annals of Math. Stat.*, Vol. 14 (1943), pp. 333-362.

from 1, $\cdots$, N without repetition have been selected from the table of random numbers. The elements associated with these integers would be a random sample. It is easy to see that the two procedures are equivalent.

A random sampling design that is not unrestricted is said to be restricted. There are many types of restricted random sampling designs of which what we call systematic designs are only one. Among these restricted designs are stratified, cluster, double, matched, polynomial, and other sampling designs, each having been developed as attempts to bring theory and practice together, to suggest improvements in practice, and to solve problems arising in practice.

By a systematic sampling design is meant a classification of the N elements into k classes, $S_1, \cdots, S_k$ where S_i consists of $x_i, x_{i+k}, \cdots, x_{i+(n-1)k}$, and a random sampling procedure for selecting one of the S_i.

It is thus clear that a systematic sampling design is a type of cluster sampling design. It will be shown that the new aspect of cluster sampling introduced in systematic sampling is that a knowledge of the order of the elements in the population is used to obtain the values of the intraclass correlation coefficient and changes in the value of that coefficient as the size of sample changes.

Sampling designs may involve combinations of random and systematic sampling designs, as well as random and nonrandom sampling procedures.

The population from which these samples are drawn may or may not be stratified and the sampling units may be single elements or clusters of elements.

4. Bases for selecting among sample designs. From the many sampling designs that can be constructed in order to obtain desired estimates, one will be chosen for use on the bases of administrative considerations, cost, and sampling error. It has become customary, on the basis of limiting distribution theory and the theory of best linear unbiased estimates to use the standard deviation of the sample estimate about the character estimated as the measure of sampling error.

Although in this paper we shall continue this practice, it must be pointed out that as more sampling designs are constructed, there is the danger that for some of these designs the limiting distribution theory is not valid, and the use of the standard error becomes more a matter of custom than the result of analysis. This danger is present for systematic sampling designs and is being further investigated.

It is perhaps desirable to remark that bias, consistency, and efficiency are properties of the sampling design and estimation functions used, not of the particular sample obtained. Any estimate based on a sample will probably differ from the character estimated. It is the function of statistical analysis to indicate how large this difference may be.

5. Notation. The letter, P, with appropriate subscripts is used for population, and subpopulations such as strata.

The number of strata is denoted by L, and the number of elements in the i^{th} stratum is denoted by N_i. Sizes of sample are denoted by n with appropriate subscripts.

The arithmetic mean of the elements of a population or subpopulation is denoted by $\bar{x}$ with appropriate subscripts.

Any particular subclass of a population as defined by the sampling design is denoted by S with subscripts. Estimates based on an S with subscripts are denoted by $\tilde{x}$ with subscripts.

6. Unstratified systematic sampling, the sampling unit consisting of one element. The values assumed by the subscripts used in this section are given in Appendix A.

Let the population, P, consist of N elements $x_1, \cdots, x_N$. It is desired to estimate the arithmetic mean, $\bar{x}$, of P.

Let[5] $N = kn$, and let the class S_i consist of the n elements $x_i, x_{i+k}, \cdots x_{i+(n-1)k}$. Then, the systematic sampling design for estimating $\bar{x}$ from a sample of size n, consists of the k classes, $S_1, \cdots, S_k$, and the requirement that the sampling procedure be such that the probability is $1/k$, that S_i is the class selected by the sampling procedure.

Let $\bar{x}_i$ be the arithmetic mean of the elements of S_i, i.e., $n\bar{x}_i = x_i + x_{i+k} + \cdots + x_{i+(n-1)k}$, and let $\tilde{x}$ be the sample mean, i.e., $\tilde{x} = \bar{x}_i$ if S_i is selected by the sampling procedure.

In dealing with systematic sampling, we shall have occasion to use both the circular and non-circular definitions of the serial correlation coefficients and the associated serial variances.

We shall assume that if $h > kn$ then $x_h = x_{h-kn}$. This is used in the circular definitions.

Let
$$kn\sigma^2 = \sum_{\nu}(x_\nu - \bar{x})^2,$$
and let
$$knC_{k\mu} = \sum_{\nu}(x_\nu - \bar{x})(x_{\nu+k\mu} - \bar{x}).$$

Then, the circular definition of the serial correlation coefficient $\rho_{k\mu}$ is $\sigma^2\rho_{k\mu} = C_{k\mu}$, which we shall use unless n is even, when we define $\rho_{kn/2}$ by the equation

$$2\sigma^2\rho_{kn/2} = C_{kn/2},$$

in order to simplify the writing of the formula for $\sigma_{\tilde{x}}^2$.

Similarly, if we define the serial variance, $s_{k\mu}$, by the equation $kns_{k\mu} = \sum_{\nu}(x_\nu - x_{\nu+k\mu})^2$, then we are using the circular definition of the serial variance. The circular definition of the serial variance ratio $v_{k\mu}$ is then $\sigma^2 v_{k\mu} = s_{k\mu}$ which we shall use unless n is even, when we define $v_{kn/2}$ by the equation

$$2\sigma^2 v_{kn/2} = s_{kn/2}.$$

The non-circular definitions of the serial correlations and serial variances are given by

$$(1)\qquad \begin{aligned} k(n - \delta)C'_{k\delta} &= \sum_j (x_j - \bar{x})(x_{j+k\delta} - \bar{x}), \\ \sigma^2 \rho'_{k\delta} &= C'_{k\delta}, \\ k(n - \delta)s'_{k\delta} &= \sum_j (x_j - x_{j+k\delta})^2, \end{aligned}$$

and

$$\sigma^2 v'_{k\delta} = s'_{k\delta}.$$

The intraclass correlation coefficient $\bar{\rho}_k$ is defined by the equation

$$\sigma^2 \bar{\rho}_k = \mathcal{E}(x_\mu - \bar{x})(x_\nu - \bar{x}),$$

where the random process consists in first sampling one of the S_i at random and then selecting two of the x's at random from the S_i that was selected. Then, since

$$k\sigma^2_{\bar{x}} = \sum_i (\bar{x}_i - \bar{x})^2,$$

and,

$$(2)\qquad \sigma^2 \bar{\rho}_k = (n/n - 1)\sigma^2_{\bar{x}} - (1/n - 1)\sigma^2$$

we have

$$(3)\qquad \sigma^2_{\bar{x}} = \frac{1}{n}(1 + (n - 1)\bar{\rho}_k)$$

It is easy to see from (1) that the intraclass correlation coefficient is given by

$$\begin{aligned} \bar{\rho}_k &= \frac{2}{n(n - 1)} \sum_\delta (n - \delta)\rho'_{k\delta} \\ &= \frac{2}{n - 1} \sum_\mu \rho_{k\mu}, \end{aligned}$$

and that consequently, if n is odd, $\bar{\rho}_k$ is the arithmetic mean of the $\rho_{k\mu}$ while if n is even, $\bar{\rho}_k$ is equal to the arithmetic mean of the $\rho_{k\mu}$ multiplied by $n/(n - 1)$.

THEOREM[10]: *Using the systematic sampling design, the estimate $\tilde{x}$ is an unbiased estimate of $\bar{x}$, and has variance $\sigma^2_{\bar{x}}$ where*

$$(4)\qquad \begin{aligned} \sigma^2_{\bar{x}} &= \sigma^2 \left\{1 - \frac{1}{n^2} \sum_\delta (n - \delta)v'_{k\delta}\right\} \\ &= \sigma^2 \left(1 - \frac{1}{n} \sum_\mu v_{k\mu}\right) \\ &= \frac{\sigma^2}{n} \left\{1 + \frac{2}{n} \sum_\delta (n - \delta)\rho'_{k\delta}\right\} \\ &= \frac{\sigma^2}{n}(1 + 2\sum_\mu \rho_{k\mu}) \\ &= \frac{\sigma^2}{n}\{1 + (n - 1)\bar{\rho}_k\}. \end{aligned}$$

[10] A proof of Theorem 1 that is somewhat simpler to follow but which, in the authors opinion, is not as informative as that given below could be obtained by substituting for $\bar{\rho}_k$ using equations (2) and (3). The lemmas in Appendix B are, of course, of interest in themselves in finite sampling.

Proof: From the definitions of expected value, $\bar{x}_i$, $\tilde{x}$, and the systematic sampling design, it follows that $\tilde{x}$ is a variate with possible values $\bar{x}_1, \cdots, \bar{x}_k$, the probability that $\tilde{x} = \bar{x}_i$ being $1/k$. Then

$$k\mathfrak{E}\tilde{x} = \bar{x}_1 + \cdots + \bar{x}_k, \tag{5}$$

and, when the values of the $\bar{x}_i$ are substituted in (5), it follows that $\mathfrak{E}\tilde{x} = \bar{x}$, that is, $\tilde{x}$ is an unbiased estimate of $\bar{x}$.

Having calculated $\mathfrak{E}\tilde{x}$, it is necessary to calculate $\mathfrak{E}\tilde{x}^2$ in order to evaluate $\sigma_{\tilde{x}}^2$. From the definition of expected values, it follows that

$$k\mathfrak{E}\tilde{x}^2 = \bar{x}_1^2 + \cdots + \bar{x}_k^2, \tag{6}$$

and when the values of the $\bar{x}_i$ are substituted in (6), it follows that

$$n^2k\mathfrak{E}\tilde{x}^2 = \sum_{i,\alpha,\gamma} x_{i+(\alpha-1)k}\, x_{i+(\gamma-1)k} \tag{7}$$

Then, when $f(u)$ is replaced by u in Lemma 6 of Appendix B, it follows, from the definition of the variance, that $\sigma_{\tilde{x}}^2 = \left(\frac{1}{kn}\right)\sum_{\nu} (x_\nu - \bar{x})^2 - \left(\frac{1}{kn^2}\right)\sum_{\delta,j} (x_j - x_{j+k\delta})^2 = \sigma^2 - \frac{1}{n^2}\sum_{\delta}(n - \delta)s'_{k\delta}$, and when $f(u)$ is replaced by u in Lemma 8, it follows that

$$\begin{aligned}\sigma_{\tilde{x}}^2 &= \left(\frac{1}{kn^2}\right)\sum_{\nu} (x_\nu - \bar{x})^2 + \left(\frac{2}{kn^2}\right)\sum_{\delta,j} (x_j - \bar{x})(x_{j+k\delta} - \bar{x}) \\ &= \left(\frac{1}{n}\right)\sigma^2 + \frac{2}{n^2}\sum_{\delta} (n - \delta)c'_{k\delta}.\end{aligned}$$

If in Lemma 9 of Appendix B we now replace $f(x_j, x_{j+k\delta})$ by $(x_j - x_{j+k\delta})^2$ then $\sigma_{\tilde{x}}^2 = \sigma^2 - \left(\frac{1}{n}\right)\sum_{\mu} s_{k\mu}$,

and if we replace $f(x_j, x_{j+k\delta})$ by $(x_j - \bar{x})(x_{j+k\delta} - \bar{x})$ then

$$\sigma_{\tilde{x}}^2 = \frac{1}{n}\left(\sigma^2 + 2\sum_{\mu} c_{k\mu}\right).$$

Finally, we have, then

$$\sigma_{\tilde{x}}^2 = \sigma^2\left(1 - \frac{1}{n}\sum_{\mu} v_{k\mu}\right),$$

and

$$\sigma_{\tilde{x}}^2 = \frac{\sigma^2}{n}\left(1 + 2\sum_{\mu} \rho_{k\mu}\right).$$

7. Possible values of the $\rho'_{k\delta}$, $\rho_{k\mu}$ and $\sigma_{\tilde{x}}^2$. Let us investigate briefly the effects of different patterns of variation on the values of $\rho'_{k\delta}$ and $\sigma_{\tilde{x}}^2$. Now $\sigma^2\rho'_{k\delta} = \frac{1}{k(n - \delta)}\sum_{j} (x_j - \bar{x})(x_{j+k\delta} - \bar{x})$. Suppose that $x_i = x_{i+k\delta}$, $\delta = 1, \cdots, n - 1$,

$i = 1, \cdots, k$. Then $\sum_{\nu} (x_\nu - \bar{x})^2 = n \sum_i (x_i - \bar{x})^2$, and $\sum_j (x_j - \bar{x})(x_{j+k\delta} - \bar{x}) = (n - \delta) \sum_i (x_i - \bar{x})^2$. Upon substitution it follows that $\rho'_{k\delta} = 1$, and $\sigma_{\bar{x}}^2 = \sigma^2$. This result for $\sigma_{\bar{x}}^2$ is intuitively clear, since all the variability is among the possible samples, and thus any particular systematic sample is equivalent to one observation.

Suppose, on the other hand, that $x_{k\delta+\alpha} = x_{k\delta+\beta}\,\alpha, \beta = 1, \cdots, k; \delta = 1, \cdots, n - 1$. Then $\sum_{\nu} (x_\nu - \bar{x})^2 = k \sum_{\alpha} (x_{i+(\alpha-1)k} - \bar{x})^2$ for any i, $i = 1, \cdots, k$, and $\sum_j (x_j - \bar{x})(x_{j+k\delta} - \bar{x}) = k \sum_{\lambda} (x_{i+(\lambda-1)k} - \bar{x})(x_{i+(\lambda+\delta-1)k} - \bar{x})$. Furthermore $0 = [\sum_{\alpha} (x_{i+(\alpha-1)k} - \bar{x})]^2 = \sum_{\alpha} (x_{i+(\alpha-1)k} - \bar{x})^2 + 2 \sum_{\lambda,\delta} (x_{i+(\lambda-1)k} - \bar{x})(x_{i+(\lambda+\delta-1)k} - \bar{x})$. Hence

$$2\sum_{\delta} \frac{n - \delta}{n} \rho'_{k\delta} = -1 \quad \text{and} \quad \sigma_{\bar{x}}^2 = 0.$$

It is possible to construct examples in which any particular $\rho'_{k\delta} = -1$, but in such cases the remaining $\rho'_{k\delta}$ each vanish. It is well known that the minimum value of $\bar{\rho}_k$ is $-1/(n - 1)$.

Finally, let us consider the expected values of $\rho'_{k\delta}$ and $\sigma_{\bar{x}}^2$ if the x's have been assigned their subscripts at random. These values are $\rho'_{k\delta} = -1/(nk - 1)$ and $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{nk - n}{nk - 1}\right)$.

In most practical applications of systematic sampling it will be highly unlikely that the distribution of the x's will be such that the x's may be said to have been assigned their subscripts at random. In general, there will be logical reasons to expect that the x's will have some fundamental trend. Thus, information will often be available, or may be obtained by a small subsample, on the basis of which a decision can be made to use some approach differing from that of assuming the subscripts of the x's to have been assigned at random.

8. Estimates of the parameters.

The formulae obtained in section 6 for the variance of the mean of a systematic sample are population formulae. Their values depend on the values of all the elements of the population. However, even in tests of possible sampling procedures, we rarely have available the resources with which to study the entire population. Consequently, it becomes necessary to investigate the possibility of estimating the population variances and serial correlations from samples. It will be shown that the estimates of the variances and correlations derived from a single S_i are biased and inconsistent whereas it will be possible to construct unbiased or consistent estimates from samples of more than one of the S_i. The sampling variations of these estimates must be left for further study.

Let us assume that instead of sampling only one of the S_i, as we did in section

6, we sampled g of the S_i at random. Then our sample would consist of all the elements in the S_β. The sample mean, $\hat{x}$, is defined by

$$g\hat{x} = \sum_\beta \bar{x}_\beta$$

if the subscripts of our sample classes are $i_1, \cdots, i_g$.

Then it is easy to see that $\hat{x}$ is unbiased Furthermore, since we can regard this sampling procedure as the sampling of g of k elements at random, it follows that $\sigma_{\hat{x}}^2 = \frac{k-g}{k-1}\frac{1}{g}\,\sigma_{\bar{x}}^2$ and, we have evaluated $\sigma_{\bar{x}}^2$ in section 6.[11]

Since

$$k\sigma_{\bar{x}}^2 = \sum_i (\bar{x}_i - \bar{x})^2,$$

we shall consider estimating $\sigma_{\bar{x}}^2$ by $s_{\bar{g}}^2$ where

$$gs_{\bar{g}}^2 = \sum_\beta (\bar{x}_\beta - \hat{x})^2.$$

Now, since $E\,\hat{x}^2 = \sigma_{\hat{x}}^2 + \bar{x}^2$, and

$$E\sum_\beta \bar{x}_\beta^2 = \frac{g}{k}\sum_i \bar{x}_i^2 = g(\sigma_{\bar{x}}^2 + \bar{x}^2)$$

it follows that $\mathcal{E}s_{\bar{g}}^2 = \sigma_{\bar{x}}^2$, and hence $s_{\bar{g}}^2$ is an unbiased estimate of $\sigma_{\bar{x}}^2$. Furthermore $\mathcal{E}s_{\hat{g}}^2 = \frac{g(k-1)}{k-g}\,\sigma_{\hat{x}}^2$.

We now turn to estimates of the $\rho_{k\mu}$ and σ^2.

Let

$$gn\hat{s}_g^2 = \sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \hat{x})^2$$

and let

$$gn\hat{c}_{k\mu g} = \sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \hat{x})(x_{\beta+(\alpha+\mu-1)k} - \hat{x}).$$

Then, it may be shown that

$$\mathcal{E}\hat{s}_g^2 = \sigma^2 - \sigma_{\hat{x}}^2,$$

and

$$\mathcal{E}\hat{c}_{k\mu g} = C_{k\mu} - \sigma_{\hat{x}}^2.$$

Hence

$$\mathcal{E}\hat{s}_g^2 + s_{\bar{g}}^2\left(\frac{k-g}{g(k-1)}\right) = \sigma^2$$

[11] It may be wondered why we sample these S_i at random rather than systematically. If we sampled the S_i systematically, it would be equivalent to taking a single systematic sample having smaller intervals between elements of the sample. Furthermore, we could not derive the unbiased estimates of the sampling variance that we can now.

and

$$\mathcal{E}\hat{c}_{k\mu g} + s_{\hat{g}}^2\left(\frac{k-g}{g(k-1)}\right) = C_{k\mu} .$$

The estimate, $r_{k\mu}$, of $\rho_{k\mu}$ defined by

$$\hat{c}_{k\mu g} + s_{\hat{g}}^2\left(\frac{k-g}{g(k-1)}\right) = r_{k\mu}\left[\hat{s}_g^2 + s_{\hat{g}}^2\left(\frac{k-g}{g(k-1)}\right)\right]$$

is thus biased but in many cases the bias will be small Of course, if $\mu = \frac{n}{2}$ when n is even then $r_{kn/2}$ is multiplied by 2 to estimate $\rho_{kn/2}$ as previously defined. Another approach would be to consider

$$gn\,{}_w\hat{s}_g^2 = \sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \bar{x}_\beta)^2,$$

and

$$gn\,{}_w\hat{c}_{k\mu g} = \sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \bar{x}_\beta)(x_{\beta+(\alpha+\mu-1)k} - \bar{x}_\beta).$$

When that is done, it follows that

$$\mathcal{E}\,{}_w\hat{s}_g^2 = \sigma^2 - \sigma_{\bar{\bar{x}}}^2 ,$$

and

$$\mathcal{E}\,{}_w\hat{c}_{k\mu g} = C_{k\mu} - \sigma_{\bar{\bar{x}}}^2 ,$$

and

$$\mathcal{E}({}_w\hat{s}_g^2 + s_{\hat{g}}^2) = \sigma^2,$$

$$\mathcal{E}({}_w\hat{c}_{k\mu g} + s_{\hat{g}}^2) = C_{k\mu} .$$

Another estimate of $\rho_{k\mu}$ is thus defined by the equation

$${}_w\hat{c}_{k\mu g} + s_{\hat{g}}^2 = {}_wr_{k\mu}[{}_w\hat{s}_g^2 + s_{\hat{g}}].$$

When $g = 1$, $s_{\hat{g}}^2 = 0$ and we are unable to provide unbiased estimates of σ^2, $C_{k\mu}$, and $\sigma_{\bar{\bar{x}}}^2$ from the sample. However, since

$$\frac{1 - r_{k\mu}}{1 - r_{k\mu'}} = \frac{\hat{s}_g^2 - \hat{c}_{k\mu g}}{\hat{s}_g^2 - \hat{c}_{k\mu' g}},$$

it follows that approximately

$$\frac{1 - \rho_{k\mu}}{1 - \rho_{k\mu'}} = \mathcal{E}\,\frac{1 - r_{k\mu}}{1 - r_{k\mu'}},$$

since $\mathcal{E}[\hat{s}_g^2 - \hat{c}_{k\mu g}] = \sigma^2 - C_{k\mu}$. Similar equations hold for the ${}_wr_{k\mu}$.

When we estimate the $\rho'_{k\delta}$, then the "within class" definition is simpler. Let $k(n - \delta)_w c'_{k\delta g} = \sum_{i,\lambda} (x_{i+(\lambda-1)k} - {}_{1\delta}\bar{x}_i)(x_{i+(\lambda+\delta-1)k} - {}_{2\delta}\bar{x}_i)$, where

$$(n - \delta)_{1\delta}\bar{x}_i = \sum_{\lambda} x_{i+(\lambda-1)k} ,$$

$$(n - \delta)_{2\delta}\bar{x}_i = \sum_{\lambda} x_{i+(\lambda+\delta-1)k} ,$$

and let

$$g(n - \delta)_w \hat{c}'_{k\delta g} = \sum_{\beta,\lambda} (x_{\beta+(\lambda-1)k} - {}_{1\delta}\bar{x}_\beta)(x_{\beta+(\lambda+\delta-1)k} - {}_{2\delta}\bar{x}_\beta).$$

Let

$$k_b c'_{k\delta g} = \sum_{i} ({}_{1\delta}\bar{x}_i - \bar{x})({}_{2\delta}\bar{x}_i - \bar{x}),$$

and let

$$g_b \hat{c}_{k'}, = \sum_{\beta} ({}_{1\delta}\bar{x}_\beta - \bar{x})({}_{2\delta}\bar{x}_\beta - \bar{x}),$$

Then

$$c'_{k\delta} = {}_w c'_{k\delta g} + {}_b c'_{k\delta g} ,$$

and

$$\mathcal{E}({}_w\hat{c}'_{k\delta g} + {}_b\hat{c}'_{k\delta g}) = C'_{k\delta} .$$

Thus, as estimates of the $\rho'_{k\delta}$ we obtain $r'_{k\delta}$ where

$${}_w\hat{c}'_{k\delta g} + {}_b\hat{c}'_{k\delta g} = r'_{k\delta}({}_w s_g^2 + s_g^2).$$

In cases where $\bar{x}$ is known simpler estimates of the $\rho_{k\mu}$, $\rho'_{k\delta}$, and σ^2 may be easily obtained since

$$\mathcal{E}\sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \bar{x})(x_{\beta+(\alpha+\mu-1)k} - \bar{x}) = gnC_{k\mu} ,$$

$$\mathcal{E}\sum_{\beta,\lambda} (x_{\beta+(\lambda-1)k} - \bar{x})(x_{\beta+(\lambda+\delta-1)k} - \bar{x}) = g(n-\delta)C'_{k\delta} ,$$

and

$$\mathcal{E}\sum_{\beta,\alpha} (x_{\beta+(\alpha-1)k} - \bar{x})^2 = gn\sigma^2.$$

Thus, in pilot studies, when $\bar{x}$ is known it is possible to estimate the parameters in $\sigma_{\bar{x}}^2$ from even a single sample.

9. Changes in the variance with changing size of sample. The chief reasons for expressing the variance of a systematic sampling design in terms of the variance of a random sample and the serial correlation coefficients were

1. To enable the making of comparisons with random and other sampling designs

2. To simplify the analysis of causes for the difference in the efficiencies of the systematic and random designs, and
3. To simplify the making of estimates of the variance for different sizes of sample.

In this section we are concerned with the third of these reasons. We shall discuss only the $\rho_{k\mu}$ since the analysis in terms of the $\rho'_{k\delta}$ is very similar.

The problem with which we are concerned is the estimation of the function, $\bar{\rho}_k$, of k. In order to show how this may be done for all values of k when the $\rho_{k\mu}$ have been computed for one value of k, let us first note that since σ^2 does not depend on k we may confine our considerations to the $C_{k\mu}$. In section 6 we have defined $C_{k\mu}$ by the equation

$$knC_{k\mu} = \sum_{\nu} (x_\nu - \bar{x})(x_{\nu+k\mu} - \bar{x}).$$

Thus, if we wish to evaluate the $C_{k'\mu'}$ where k' is such that $k' \neq k$ and $k'n' = kn = N$, we have the result $C_{k'\mu'} = C_{k\mu}$ if $k'\mu' = k\mu$ and, thus, for any given values k' and μ', we have

$$C_{k'\mu'} = C_{k\ k'\mu'/k}$$

where we have replaced μ by $\dfrac{k'\mu'}{k}$.

This procedure will involve, if $k' < k$, some interpolation, but if the $\rho_{k\mu}$ are plotted against μ, this interpolation may often be carried through graphically. However, it is usually advisable to take k so that the possible values of k' are such that $k' > k$.

In some cases it may be possible to construct a correlation function. For example, if the x_ν may be represented by a polynomial in ν, then $\rho'_{k\delta}$ may be represented by a polynomial in δ. From that fact we conclude that if the x_ν vary about a smooth trend the $\rho'_{k\delta}$ will also vary about a smooth trend and it may be possible to interpolate. Further investigation of this problem is necessary.

10. Stratified systematic sampling. In sampling practice it is customary to deal with stratified populations. The variance of an estimate based on a stratified population will usually not include the variability among the strata. Consequently, when a population is well stratified the variability of estimates based in a sample of size n will usually be considerably less than the variability of an estimate based on a random sample of size n, ignoring the strata. We now discuss the theory of systematic sampling from a stratified population.

Let us assume that the population, P, consists of L strata, $P_1, \cdots, P_L$, the ath of which contains N_a elements $x_{a1}, \cdots, x_{aN_a}$. It is desired to estimate the arithmetic mean, $\bar{x}$ of P. Let the arithmetic mean of P_a be denoted by $\bar{x}_a$. Let $N_a = k_a n_a$.

We shall consider two possible cases, the first of which is often used because of the administrative simplicity of giving identical operating instructions to the people selecting samples in different places. The results of this section will indicate when this method may be used.

Sampling Procedure I—Suppose that $k_1 = k_2 = \cdots k_L = k$, and that the sampling procedure consists in selecting one of the integers, $1, \cdots, k$ at random, each integer having a probability $1/k$ of being selected. Then, if the integer selected is, for example, i, the sample of P_a consists of $x_{ai}, x_{ai+k}, \cdots, x_{ai+(n_a-1)k}$. Thus, there are exactly k possible samples, $S_1, \cdots, S_k$, each having probability $1/k$ of being the actual sample obtained by performing the sampling procedure.

Sampling Procedure II—The sampling procedure consists in selecting one of the integers $1, \cdots, k_a$ at random, for each value of a, each integer having a probability of $1/k_a$ of being selected. Then, there are exactly $k_1 \cdot \cdots \cdot k_L$ possible samples, each having probability $1/k_1 \cdot \cdots \cdot k_L$ of being the actual sample obtained by performing the sampling procedure.

Other sampling procedures for stratified sampling, of course, exist. The two listed above, however, cover most practical problems except those involving cluster sampling. These will be treated in a later paper. Furthermore, from the conclusions derived concerning these procedures it will be possible to infer conclusions concerning other stratified sampling procedures.

Let S_{ai} be the class of elements $x_{ai}, x_{ai+k}, \cdots, x_{ai+(n_a-1)k}$. We consider sampling procedure I. A systematic sample of size n_a is to be selected from P_a. The possible samples are $S_1, \cdots, S_k$ where S_i consists of all the elements in $S_{1i}, \cdots, S_{Li}$. Let the arithmetic mean of the elements in S_{ai} be denoted by $\bar{\bar{x}}_{ai}$. Let the arithmetic mean of the sample from P_a be denoted by $\bar{\bar{x}}_a$ and let the sample mean be denoted by $\bar{\bar{x}}$, where

$$N\bar{\bar{x}} = N_1\bar{\bar{x}}_1 + \cdots + N_L\bar{\bar{x}}_L .$$

Then $N\mathcal{E}\bar{\bar{x}} = \sum_a N_a\mathcal{E}\bar{\bar{x}}_a = \sum_a N_a \frac{1}{k} \sum_i \bar{\bar{x}}_{ai} = N\bar{\bar{x}}.$

It follows from Appendix C, that

$$\sigma^2_{\bar{\bar{x}}} = \frac{1}{N^2} \sum_{a,b} N_a N_b \, \sigma_{\bar{\bar{x}}_a \bar{\bar{x}}_b}$$

where

$$\begin{aligned}\sigma_{\bar{\bar{x}}_a \bar{\bar{x}}_b} &= \mathcal{E}(\bar{\bar{x}}_a - \bar{x}_a)(\bar{\bar{x}}_b - \bar{x}_b)\\ &= \frac{1}{k} \sum_i (\bar{\bar{x}}_{ai} - \bar{x}_a)(\bar{\bar{x}}_{bi} - \bar{x}_b).\end{aligned}$$

Although the expression for $\sigma_{\bar{\bar{x}}_a \bar{\bar{x}}_b}$ can be further simplified, the important fact is that if corresponding items in different strata are positively correlated, it is inadvisable to use sampling procedure I unless other considerations than sampling error are dominant. But if the corresponding items are negatively correlated then sampling procedure I will yield a smaller variance than sampling procedure II.

We now consider sampling procedure II. The difference between sampling procedures I and II is that in sampling procedure II we know that $\sigma_{\bar{x}_a \bar{x}_b} = 0$, if $a \neq b$ because of the separate selection of sample in each stratum. Thus, under sampling procedure II, $\sigma_{\bar{x}}^2 = \frac{1}{N^2} \sum_a N^2 \sigma_{\bar{x}_a}^2$ where $\sigma_{\bar{x}_a}^2$ has been derived in section 6.

11. A comparison of the efficiences of systematic and random sampling procedures. The study of any sampling technique is incomplete unless some comparisons are made with other possible sampling techniques. In this section the systematic sampling procedure is compared with the unrestricted random and stratified random sampling procedure.

The means and variances associated with the random and stratified random sampling procedures will be denoted by the use of primes (′) and double primes (″) respectively.

Then we know that

$$\sigma_{\bar{x}}^2 / \sigma_{\bar{x}'}^2 = (1 + 2 \sum_\mu \rho_{k\mu}) \left(\frac{kn - 1}{kn - n} \right)$$

and consequently $\sigma_{\bar{x}}^2 < \sigma_{\bar{x}'}^2$ if

$$\sum_\mu \rho_{k\mu} < -(n - 1)/2(kn - 1).$$

If n is large relative to k, we may use $-\frac{1}{2}k$ as an approximation to $-(n - 1)/2(kn - 1)$.

In order to make more specific comparisons, it is useful to assume that the population elements x_ν are given by some function of ν, and to assume some functions such as

$$x_\nu = A_0 + A_1 \nu + \cdots + A_h \nu^h,$$

or

$$\begin{aligned} x_\nu = B_0 + A_1 \sin \frac{2\pi\nu}{N} + B_1 \cos \frac{2\pi\nu}{N} \\ + \cdots \\ + A_h \sin \frac{2\pi h\nu}{N} + B_h \cos \frac{2\pi h\nu}{N}, \end{aligned}$$

and then to investigate the efficiencies of the various possible sampling procedures on the bases of such assumed distributions of the x_ν. It should be noted that the use of the systematic sampling technique involves the assumption that it is possible to order the elements of the population in a logical way, and then use this ordering in selecting the sample systematically.

We shall now consider several possibilities. Let us first note that if we are sampling but one element from a stratum then the variance of the stratum sample mean is the same whether the sampling is random or systematic. On the other hand, it follows from section 10 that if we stratify the population into L strata so that a systematic sample of size L chooses the jth element of each stratum, say, then the variance of the mean of the stratified random sample will be greater or less than the variance of the mean of the systematic sample depending on whether the average correlation between strata sample means in the systematic sample is negative or positive.

Let us now consider the origin of the warnings against the use of systematic samples from a population having a periodic distribution. If k is the period, the correlation between the strata means of the systematic sample is $+1$ and hence the random sample is superior. However, if the period is $2k$ then we shall show that the systematic sample will probably have a smaller variance.

Suppose that the period is $2k$ and that within two adjoining strata of size k we always have $x_1 = x_{2k}$, $x_2 = x_{2k-1}$, $\cdots$, $x_k = x_{k+1}$ and $x_i - \bar{x} = -(x_{k+i} - \bar{x})$. Then, if we are sampling one element from each stratum, the correlation between the systematic sample means, (the individual elements in this case), will be -1 if the strata subscripts differ by an odd number and $+1$ if the strata subscripts differ by an even number.

The variance within each of the n strata is σ_1^2, where

$$k\sigma_1^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2.$$

The variance between strata means is zero. Hence $\sigma^2 = \sigma_1^2$. The variance of the mean of an unrestricted random sample of size n where $n = L$ is then $\sigma_{\bar{x}'}^2 = \frac{N - L}{N - 1}\frac{\sigma_1^2}{L}$ and the variance of the stratified random sampling mean is $\sigma_{\bar{x}''}^2 = (1/L)\sigma_1^2$ while the variance of the systematic sampling mean is

$$\sigma_{\bar{x}}^2 = \frac{\sigma_1^2}{L^2} \sum_{i,j=1}^{L} (-1)^{i-j}(2 - \delta_{ij})$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

Then it may be shown that if L is even $\sigma^2 = 0$ while, if L is odd $\sigma^2 = (1/L^2)\sigma_1^2$.

Consequently, the efficiency of the systematic sample mean is greater than the efficiency of the stratified random sample mean if the population has a periodic distribution and the size of stratum is half the period. It should be noted that the same situation holds for k equal to an even or odd multiple of half the period as held for k equal to the period of half the period.

The situation is quite different if we assume that the elements of the population have a straight line distribution. Without loss of generality, we may assume that the straight line distribution is given by $x_\nu = \nu$. Then for an unrestricted random sample of size n, the sample mean being denoted by $\bar{x}'$ we have $\mathcal{E}\bar{x}' = \bar{x} = \frac{1}{2}(kn + 1)$,

$$\sigma^2 = \frac{k^2n^2 - 1}{12},$$

and

$$\sigma^2_{\bar{x}'} = \frac{(k-1)(kn+1)}{12}.$$

For a stratified random sampling design, let us assume that $N_1 = \cdots = N_L = \frac{cN}{n}$ where c may equal any of the integers $1, 2, \cdots, n$; i.e. $L = \frac{n}{c}$. Let $n_1 = \cdots = n_L = c$. Then $\sigma^2_{\bar{x}''} = \frac{c^2}{n^2}\frac{k-1}{ck-1}\sum_a \sigma^2_a$ where $\bar{x}''$ is the sample mean of the stratified random sample. If the ath stratum contains $x_{(a-1)\frac{cN}{n}+1}, \cdots, x_{a\frac{cN}{n}}$ then $\sigma^2_a = \frac{c^2k^2-1}{12}$ and $\sigma^2_{\bar{x}''} = \frac{c}{n}\cdot\frac{k-1}{ck-1}\frac{c^2k^2-1}{12}$. Finally

$$\begin{aligned}\sigma^2_{\bar{x}} &= \sigma^2 - \frac{1}{kn^2}\sum_\delta\sum_j (x_j - x_{j+k\delta})^2\\ &= \sigma^2 - \frac{k^2(n^2-1)}{12}\\ &= \frac{k^2-1}{12}.\end{aligned}$$

To summarize

$$\begin{aligned}\sigma^2_{\bar{x}'} &= \frac{(k-1)(kn+1)}{12} = \frac{(k-1)(N+1)}{12},\\ \sigma^2_{\bar{x}''} &= \frac{c(k-1)(ck+1)}{12n} = \frac{(k-1)\left(\frac{N}{L}+1\right)}{12L},\\ \sigma^2_{\bar{x}} &= \frac{k^2-1}{12}.\end{aligned}$$

It is clear that both $\sigma^2_{\bar{x}}$ and $\sigma^2_{\bar{x}''}$ are less than $\sigma^2_{\bar{x}'}$. However $\sigma^2_{\bar{x}}/\sigma^2_{\bar{x}''} = \frac{L(k+1)}{\frac{N}{L}+1}$.

Since $kn = N$ and $cL = n$ it follows that $k = \frac{N}{cL}$ and hence $\sigma^2_{\bar{x}} < \sigma^2_{\bar{x}''}$ if $N > L(L-1)$ and $c \geq \frac{L}{1 - \frac{L(L-1)}{N}}$. In all cases $\frac{N}{L} \geq c$. It follows therefore that for a value of c to exist we must have $\frac{N}{L} > \frac{NL}{N - L(L-1)}$ as a result of which we find that N must exceed $2L^2 - L$. Hence $\sigma^2_{\bar{x}} \leq \sigma^2_{\bar{x}''}$ if $N > 2L^2 - L$ and $c \geq \frac{L}{1 - \frac{L(L-1)}{N}}$. Otherwise $\sigma^2_{\bar{x}} > \sigma^2_{\bar{x}''}$.

This result follows from two facts:

(1) If one element is being taken from each stratum then the high average correlation between strata means results in the efficiency of the stratified random sampling mean being greater than the efficiency of the systematic sampling mean, despite the equal within stratum variances.

(2) If more than one element is being taken from each stratum then the within stratum variance of the systematic sampling mean is less than the within stratum variance of the stratified random sampling mean and if the size of stratum and sample from stratum are large enough, the smaller within stratum variance of the systematic sample more than compensates for the correlation among strata means.

Of course, in a straight line distribution there are much more efficient methods of defining a stratified random sample than that we have used. Furthermore, more efficient sampling procedures than those discussed are available. However, this example will be of use in indicating the general problems that arise as well as the procedures that may be followed in attacking them.

Another comparison of systematic and stratified random sampling may be obtained by considering the x_ν to be composed of two elements, a trend function and a periodic function so that the deviations from the trend constitute a periodic function.

Let $x_\nu = \varphi_1(\nu) + \varphi_2(\nu)$, where $\varphi_1(\nu)$ is a trend function and $\varphi_2(\nu)$ is a periodic function of period $2h$, $N = 2hQ$.

Let $\varphi_2(\nu) = y_\nu$. Then $y_j = y_{2h+j} = \cdots = y_{2h(Q-1)+j}$, $j = 1, \cdots, 2h$ and $y_{2ha+j} - \bar{y} = -(y_{2ha+h+j} - \bar{y})$, $j = 1, \cdots, h$, $a = 0, \cdots, Q - 1$.

Since the sizes of sample that we shall consider for purposes of this comparison are all multiples of h we shall calculate our variances and covariances so that we obtain all the necessary information at once.

Let the mean of $\varphi_1(\nu)$ be denoted by $\bar{\varphi}_1$ and let the mean of $\varphi_2(\nu)$ be denoted by $\bar{y}$. Then $\bar{x} = \bar{\varphi}_1 + \bar{y}$ and

$$\begin{aligned}
N\sigma^2 &= \sum_\nu (x_\nu - \bar{x})^2 \\
&= \sum_\nu [\varphi_1(\nu) - \bar{\varphi}_1]^2 + \sum_\nu (y_\nu - \bar{y})^2 + 2\sum_\nu [\varphi_1(\nu) - \bar{\varphi}_1](y_\nu - \bar{y}) \\
&= \sum_{a,i} (\varphi_{1(a-1)h+i} - \bar{\varphi}_{1a})^2 + h\sum_a (\bar{\varphi}_{1a} - \bar{\varphi}_1)^2 \\
&+ \sum_{a,i} (y_{(a-1)h+i} - \bar{y}_a)^2 + h\sum_a (\bar{y}_a - \bar{y})^2 \\
&+ \sum_{a,i} [\varphi_{1(a-1)h+i} - \bar{\varphi}_{1a}](y_{(a-1)h+i} - \bar{y}_a) + h\sum_a (\bar{\varphi}_{1a} - \bar{\varphi}_1)(\bar{y}_a - \bar{y})
\end{aligned}$$

where $a = 1, \cdots, 2Q$; $i = 1, \cdots, h$, $\bar{\varphi}_{1a}$ is the arithmetic mean of $\varphi_{1[(a-1)h+1]}$, $\cdots$, $\varphi_{1[ah]}$ and $\bar{y}_a$ is the arithmetic mean of $y_{(a-1)h+1}, \cdots, y_{ah}$.

It follows from the assumptions with respect to the y_ν that $\bar{y}_a = \bar{y}$ and that $\sum_i (y_i - \bar{y}_a)^2$ is the same for each value of a and is equal to $\sum_{i=1}^{h} (y_i - \bar{y})^2$.

Since $y_i = y_{2h+i} = \cdots = y_{2h(Q-1)+i}$ and $y_{h+i} = y_{3h+i} = \cdots = y_{2h(Q-1)+h+i}$ we have

$$\sum_{\nu} [\varphi_1(\nu) - \bar{\varphi}_1](y_\nu - \bar{y}) = \sum_{i=1}^{h} (y_i - \bar{y}) \sum_{a=1}^{Q} [\varphi_{1(2a-2)h+i} - \bar{\varphi}_{12a-2}]$$

$$+ \sum_{i=h+1}^{2h} (y_i - \bar{y}) \sum_{a=1}^{Q} [\varphi_{1((2a-2)h+i)} - \bar{\varphi}_{12a-2}].$$

Since $y_i - \bar{y} = (y_{i+h} - \bar{y})$, we also have $\sum_{\nu} [\varphi_1(\nu) - \bar{\varphi}_1](y_\nu - \bar{y}) = \sum_{i=1}^{h} (y_i - \bar{y}) \left\{ \sum_{a=1}^{Q} [\varphi_1((2a - 2)h + i) - \varphi_1((2a - 1)h + i) - \bar{\varphi}_{12a-2} + \varphi_{12a-1}] \right\}$ which vanishes for example if $\varphi_1(\nu)$ is a straight line or if $\varphi_1(\nu)$ is a succession of straight lines each having length $2h$.

Let us now assume that $\varphi_1(\nu) = A + B\nu$. Then $\bar{\varphi}_{1a} = A + B\left[(a - 1)h + \frac{h + 1}{2}\right]$,

$$\varphi_1((a - 1)h + i) - \bar{\varphi}_{1a} = i - \frac{h + 1}{2},$$

$$\sum_{i} [\varphi_1((a - 1)h + i) - \bar{\varphi}_{1a}]^2 = \frac{B^2 h(h^2 - 1)}{12},$$

$$\sum_{a} (\bar{\varphi}_{1a} - \bar{\varphi}_1)^2 = \frac{B^2 h^2 (4Q^2 - 1)(2Q)}{12},$$

$$\bar{\varphi}_1 = A + B\,\frac{2h + 1}{2},$$

$$\sum_{\nu} (\varphi_1(\nu) - \bar{\varphi}_1)^2 = \frac{2hQB^2}{12}\,[4h^2Q^2 - 1].$$

Then $\sigma^2 = \sigma_y^2 + \frac{B^2}{12}[4h^2Q^2 - 1]$ where $h\sigma_y^2 = \sum_{i} (y_i - \bar{y})^2$, and the variance of the mean of an unrestricted random sample of size n is $\sigma^2_{\bar{y}'} = \frac{N - n}{N - 1}\frac{\sigma^2}{n}$.

Let us assume now that the size of stratum is mh where m is a factor of $2Q$, say $2Q/m = L_m$. Then the variance within each of the L_m strata is a constant, say, σ_1^2 where $\sigma_1^2 = \sigma_y^2 + \frac{B^2}{12}[h^2m^2 - 1]$ if m is even. If m is odd then L_m is even and in half the strata the within stratum variance is $\sigma_1^2 + \frac{1}{hm}\sum_{i=1}^{h} (y_i - \bar{y})\left(i - \frac{h + 1}{2}\right)$ while in the other strata, the within stratum variance is $\sigma_1^2 - \frac{1}{hm}\sum_{i=1}^{h} (y_i - \bar{y})\left(i - \frac{h + 1}{12}\right)$.

Then, if c elements are sampled at random from each of the L_m strata, it follows that

$$\sigma^2_{\bar{x}''} = \frac{1}{L_m}\left(\frac{mh - c}{mh - 1}\right)\frac{\sigma_1^2}{c}$$
$$= \frac{1}{L_m}\left(\frac{mh - c}{mh - 1}\right)\frac{1}{c}\left(\sigma_y^2 + \frac{B^2}{12}[h^2m^2 - 1]\right).$$

In order to evaluate the variance of the systematic sampling mean let us evaluate $\sum_j (x_j - x_{j+k\delta})^2 = k(n - \delta)s'_{k\delta}$. Now upon substituting for x_ν, it follows that $k(n - \delta)s'_{k\delta} = \sum_j (y_j - y_{j+k\delta})^2 - 2Bk\delta \sum_j (y_j - y_{j+k\delta}) + k(n - \delta)B^2k^2\delta^2$.

Then, if k is a multiple of h, it follows that $\sum_j (y_j - y_{j+k\delta}) = 0$. Furthermore, if k is an even multiple of h, then $y_j = y_{j+\delta k}$ and hence $\sum_j (y_j - y_{j+\delta k})^2 = 0$. Finally, if k is an odd multiple of h then, if δ is an odd number $y_j - y_{j+\delta k} = 2(y_j - \bar{y})$ while, if δ is even $y_j - y_{j+\delta k} = 0$ and hence

$$\sum_j (y_j - y_{j+\delta k})^2 = 4\sum_j (y_j - \bar{y})^2$$
$$= 4\,\frac{k(n - \delta)}{h}\sum_i (y_i - \bar{y})^2$$

if k is an odd multiple of h and δ is an odd number. Note that if k is an odd multiple of h, then n is an even number. Since

$$\sigma^2_{\bar{x}} = \sigma^2 - \frac{1}{n^2}\sum_\delta (n - \delta)s'_{k\delta}$$

it is necessary to evaluate $\sum_\delta (n - \delta)s'_{k\delta}$. Now, if k is an even multiple of h, it follows that $(n - \delta)s'_{k\delta} = (n - \delta)B^2k^2\delta^2$ and

$$\sum_\delta (n - \delta)s'_{k\delta} = B^2k^2\{n\sum_\delta \delta^2 - \sum_\delta \delta^3\}$$
$$= B^2k^2\,\frac{n^2(n^2 - 1)}{12}$$

Hence, if k is an even multiple of h, it follows that $\sigma^2_{\bar{x}} = \sigma^2 - \dfrac{B^2k^2(n^2 - 1)}{12}$.

On the other hand, if k is an odd multiple of h, and if δ is odd, we have $(n - \delta)s'_{k\delta} = (n - \delta)B^2k^2\delta^2 + 4(n - \delta)\sigma_y^2$ while if δ is even $(n - \delta)s'_{k\delta} = (n - \delta)B^2k^2\delta^2$.

Hence

$$\sum_\delta (n - \delta)s'_{k\delta} = \frac{B^2k^2n^2(n^2 - 1)}{12} + n^2\sigma_y^2.$$

Hence, if k is an odd multiple of h, it follows that

$$\sigma_{\bar{x}}^2 = \sigma^2 - \frac{B^2k^2(n^2 - 1)}{12} - \sigma^2 .$$

Then, if k is an even multiple of h

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \sigma_y^2 + \frac{B^2}{12}(k^2n^2 - 1) - \frac{B^2}{12}k^2(n^2 - 1)\\ &= \sigma_y^2 + \frac{B^2}{12}(k^2 - 1),\end{aligned}$$

and, if k is an odd multiple of h, then $\sigma_{\bar{x}}^2 = \dfrac{B^2}{12} B^2 (k^2 - 1)$.

Thus, systematic sampling will yield superior results if

$$c > \frac{L}{1 + \dfrac{12\sigma_y^2}{B^2(hm)(hm - 1)} - \dfrac{L(L - 1)}{N}}.$$

Since $\dfrac{N}{L} > c$ it follows that for a solution, c, to exist, we must have

$$N > 2L^2 - L - \frac{12\sigma_y^2}{B^2}\left(\frac{L^2}{N - L}\right).$$

12. Summary. In this paper we have presented the theoretical basis for systematic sampling for stratified and unstratified populations including the derivation of the variances, a study of the possible values of the parameters, estimates of the parameters, the effects of changing the size of sample, and comparisons among systematic sampling, unrestricted random sampling, and stratified random sampling. The paper contains for the case where the sampling unit consists of one element, not only the theory necessary, but in addition, some analysis of the conditions under which systematic sampling ought be used, and formulas for calculating the variances.

In later papers of this series, we shall present the theory of systematic sampling when the sampling unit is a cluster of elements, the theory when we assume we are sampling not from a finite population but an infinite population, each of whose elements is normally distributed, and further studies of various parts of the theory and practice of systematic sampling.

Appendix A

Values Assumed by Certain Variables

In order to avoid repeating the limits of summation of variables, we shall give these limits in this appendix.

TABLE I

Values Assumed by Subscripts

Letter	The letter will assume all integral values from 1 to
i	k
λ	$n - \delta$
j	$k(n - \delta)$
δ	$n - 1$
ν, ν'	kn
α	n
γ	n
μ, μ'	$n/2$ if n is even, $\frac{n-1}{2}$ if n is odd
a, b	L

The letter β will assume the values $i_1, i_2, \cdots, i_g$ where $i_1, \cdots, i_g$ are a selection of g of the k integers $1, \cdots, k$.

Appendix B

On the Limits of Some Finite Sums

The difficulties that arise in the transformation of finite sums are very similar to those that arise in the theory of transforming multiple integrals, i.e., the effects of transforming variables or order of summation on the limits of summation. Certain lemmas that have proved useful in this paper are presented separately here in a more general form.

Let $f(u)$ and $f(u, v)$ be functions of u and v that are finite for all possible values of u and v.

Lemma 1.

$$\sum_{\substack{\alpha,\gamma \\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}, x_{i+(\gamma-1)k}) = \sum_{\delta,\lambda} f(x_{i+(\lambda-1)k}, x_{i+(\lambda+\delta-1)k})$$

Proof: Let $\alpha = \lambda$ and let $\gamma = \lambda + \delta$. Since $1 \leq \alpha < \gamma$ and $\gamma \leq n$, the possible values of δ are $1, \cdots, (n - 1)$. For any fixed value of δ the possible values of λ then are 1 to $n - \delta$ since $\lambda = \gamma - \delta$ and, for a fixed value of δ the maximum value of λ is determined when $\gamma = n$. With these limits each term of f on the left side of the equation occurs once and only once on the right side of the equation. Furthermore, no additional term occurs on the right side of the equation.

Lemma 2.

$$\sum_{i} \sum_{\lambda} f(x_{i+(\lambda-1)k}, x_{i+(\lambda+\delta-1)k}) = \sum_{j} f(x_j, x_{j+k\delta}).$$

Proof: Let $j = i + (\lambda - 1)k$. Then j is a monotone increasing function of i and λ. The minimum value of j occurs when $i = 1$. In that case $j = 1$. The maximum value of j occurs when $i = k, \lambda = n - \delta$. In that case $j = (n - \delta)k$.

With these limits each term of f on the left side of the equation occurs once and only once on the right side of the equation. Furthermore, no additional term occurs on the right side of the equation.

LEMMA 3.

$$\sum_{\substack{i,\alpha,\gamma\\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}, x_{i+(\gamma-1)k}) = \sum_{\delta,j} f(x_j, x_{j+\delta k}).$$

PROOF: First apply Lemma 1 to $\sum_{\substack{\alpha,\gamma\\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}, x_{i+(\gamma-1)k})$ and then apply Lemma 2 to the resulting expression.

LEMMA 4.

$$\sum_{\delta,j} [f(x_j) + f(x_{j+k\delta})] = (n-1)\sum_{\nu} f(x_\nu).$$

PROOF: Let $m = j + k\delta$. Then for any fixed value of δ the minimum value of m occurs when $j = 1$. In that case $m = k\delta + 1$. For any fixed value of δ, the maximum value of m occurs when $j = k(n-\delta)$. In that case $m = nk$. The letter m will assume all integral values from $k\delta + 1$ to kn, and hence,

$$\sum_{\delta,j} f(x_j) + \sum_{\delta,j} f(x_{j+k\delta}) = \sum_{\delta,j} f(x_j) + \sum_{\delta,m} f(x_m).$$

If we sum δ from $n-1$ to 1 instead of from 1 to $n-1$ in $\sum_{\delta,m} f(x_m)$ we see that

$$\begin{aligned}\sum_{\delta,j} f(x_j) + \sum_{\delta,m} f(x_m) = &\sum_{j=1}^{k(n-1)} f(x_j) + \sum_{m=k(n-1)+1}^{kn} f(x_m)\\ &+ \cdots\\ &+ \sum_{j=1}^{k} f(x_j) + \sum_{m=k+1}^{kn} f(x_m)\end{aligned}$$

where the summations of x_j are terms of $\sum_{\delta,j} f(x_j)$ and the summations of x_m are terms of $\sum_{\delta,m} f(x_m)$. But $\sum_{j=1}^{k(n-\delta)} f(x_j) + \sum_{m=k(n-\delta)+1}^{kn} f(x_m) = \sum_{\nu} f(x_\nu)$ and hence Lemma 4 is proved.

LEMMA 5. *Let*

$$\sum_{i,\alpha,\gamma} f(x_{i+(\alpha-1)k}) f(x_{i+(\gamma-1)k}) = A.$$

Then

$$A = n\sum_{\nu} [f(x_\nu)]^2 - \sum_{j,\delta} [f(x_j) - f(x_{j+k\delta})]^2.$$

PROOF:

$$A = \sum_{i,\alpha} [f(x_{i+(\alpha-1)k})]^2 + 2\sum_{\substack{i,\alpha,\gamma\\ \alpha<\gamma}} f(x_{i+(\alpha-1)k}) f(x_{i+(\gamma-1)k}).$$

By Lemma 3

$$2 \sum_{\substack{i,\alpha,\gamma \\ \alpha < \gamma}} f(x_{i+(\alpha-1)k}) f(x_{i+(\gamma-1)k}) = 2 \sum_{\delta,j} f(x_j) f(x_{j+\delta k})$$

and since we have

$$2f(x_j)f(x_{j+\delta k}) = f(x_j)^2 + f(x_{j+\delta k})^2 - [f(x_j) - f(x_{j+\delta k})]^2,$$

the proof is completed by using Lemma 4.

LEMMA 6. Let $kn\bar{f} = \sum_{\nu} f(x_\nu)$. *Then*

$$A\left(\frac{1}{kn^2}\right) - \bar{f}^2 = \left(\frac{1}{kn}\right) \sum_{\nu} [f(x_\nu) - \bar{f}] - \left(\frac{1}{kn^2}\right) \sum_{j,\delta} [f(x_j) - f(x_{j+k\delta})]^2.$$

PROOF: This lemma is a direct consequence of Lemma 5.

LEMMA 7.

$$A = \sum_{\nu} [f(x_\nu) - \bar{f}]^2 + 2\sum_{j,\delta} [f(x_j) - \bar{f}][f(x_{j+k\delta}) - \bar{f}] + kn^2\bar{f}^2.$$

PROOF: From Lemma 4, it follows that

$$A = n\sum_{\nu} f(x_\nu)^2 - \sum_{j,\delta} \{[f(x_j) - \bar{f}]^2 + [f(x_{j+k\delta}) - \bar{f}]^2 + 2\sum_{j,\delta} [f(x_j) - \bar{f}][f(x_{j+k\delta}) - \bar{f}]$$

and hence, from Lemma 3, it follows that

$$A = n\sum_{\nu}[f(x_\nu) - \bar{f}]^2 + n^2k\bar{f}^2 - (n-1)\sum_{\nu} [f(x_\nu) - \bar{f}]^2$$

$$+ 2 \sum_{j,\delta}[f(x_j) - \bar{f}][f(x_{j+k\delta}) - \bar{f}],$$

whence the lemma is proved.

LEMMA 8.

$$A\left(\frac{1}{kn^2}\right) - \bar{f}^2 = \left(\frac{1}{kn^2}\right) \sum_{\nu} [f(x_\nu) - \bar{f}]^2 + \left(\frac{2}{kn^2}\right) \sum_{j,\delta} [f(x_j) - \bar{f}] [f(x_{j+k\delta}) - \bar{f}].$$

This lemma is a direct consequence of Lemma 7.

LEMMA 9. *If $h > kn$, let x_h equal x_{h-kn}. Let $f(u, v) = f(v, u)$ i.e. f is symmetric. Then, if we let*

$$d_{k\delta} = \sum_{j} f(x_j, x_{j+k\delta})$$

it follows that

$$d_{k\delta} + d_{kn-\delta} = \sum_{\nu} f(x_\nu, x_{\nu+k\delta}).$$

PROOF: Obviously

$$\sum_{\nu} f(x_\nu, x_{\nu+k\delta}) = d_{k\delta} + B,$$

where

$$B = \sum_{g=k(n-\delta)+1}^{kn} f(x_g\,, x_{g+k\delta}).$$

Now, let $h = g - (n - \delta)k$. Then

$$B = \sum_{h=1}^{\delta k} f(x_{h+(n-\delta)k}\,, x_{h+kn}).$$

Since $x_{h+kn} = x_h$, and $f(x_{h+(n-\delta)k}\,, x_h) = f(x_h\,, x_{h+(n-\delta)k})$, it follows that $B = d_{kn-\delta}$ and the lemma is proved. It is noted that the symmetry of $f(u, v)$ is necessary as well as sufficient, for if $f(x_\nu\,, x_{\nu+k\delta}) = x_\nu - x_{\nu+k\delta}$ the theorem is false.

Appendix C

Stratified Sampling

Let the population P consist of L strata $P_1\,, \cdots, P_L$. Let $\bar{x}$ be the arithmetic mean of P, and $\bar{x}_a$ the arithmetic mean of P_a. Let $\tilde{x}_a$ be the sample estimate of $\bar{x}_a$, and let $\tilde{x} = \sum_a c_a\tilde{x}_a$. Then $\mathcal{E}\tilde{x} = \sum_a c_a A_a = A$ where $\mathcal{E}\tilde{x}_a = A_a$. Let $\sigma_{\tilde{x}}^2$ be defined by $\sigma_{\tilde{x}}^2 = \mathcal{E}(\tilde{x} - \bar{x})^2$. Then $\sigma_{\tilde{x}}^2 = \mathcal{E}(\tilde{x} - A)^2 + (A - \bar{x})^2$ and hence it is easy to see that $\sigma_{\tilde{x}}^2 = \sum_{a,b} c_a c_b \sigma_{\tilde{x}_a\tilde{x}_b} + (A - \bar{x})^2$ where

$$\sigma_{\tilde{x}_a\tilde{x}_b} = \mathcal{E}(\tilde{x}_a - A_a)(\tilde{x}_b - A_b),$$

$$(A - \bar{x})^2 = \left[\sum_a \left(c_a A_a - \frac{N_a}{N}\,\bar{x}_a\right)\right]^2,$$

and if $NC_a = N_a$, then

$$(A - \bar{x})^2 = \sum_{a,b} c_a c_b (A_a - \bar{x}_a)(A_b - \bar{x}_b)$$

and $\sigma_{\tilde{x}}^2 = \sum_{a,b} c_a c_b \sigma_{\tilde{x}_a\tilde{x}_b}$

where

$$\sigma_{\tilde{x}_a\tilde{x}_b}^2 = \mathcal{E}(\tilde{x}_a - \bar{x}_a)(\tilde{x}_b - \bar{x}_b).$$

These formulae hold whatever may be the method used in sampling the ith stratum. If $\tilde{x}$ is an unbiased estimate of $\bar{x}$ and $\tilde{x}_a$ is independent of $\tilde{x}_b$, then the usual formula $\sigma_{\tilde{x}}^2 = \sum_a c_a^2 \sigma_{\tilde{x}_a}^2$ holds. The formula for $\sigma_{\tilde{x}_a\tilde{x}_b}$ will, of course, depend on whether a random, cluster, systematic, or other sampling procedure is used.

ON THE PROBABILITY THEORY OF LINKAGE IN MENDELIAN HEREDITY

By Hilda Geiringer

Bryn Mawr College

1. Introduction. If for a certain generation the distribution of genotypes is known and a certain law of heredity is assumed, the distribution of genotypes in the next generation can be computed. Suppose there are N different genotypes in the nth generation in the proportions $x_1^{(n)}, \cdots, x_N^{(n)}$ where $\sum_{i=1}^{N} x_i^{(n)} = 1$ and denote by $p^i_{\kappa\lambda}$ the probability that an offspring of two parents of types κ and λ be of type i where $\sum_{i=1}^{N} p^i_{\kappa\lambda} = 1$ for all κ and λ, and $p^i_{\kappa\lambda} = p^i_{\lambda\kappa}$. Assuming panmixia, identical distributions $x_i^{(n)}$ for males and females, etc., we can derive $x_i^{(n+1)}$ from $x_i^{(n)}$ by means of the formula

$$x_i^{(n+1)} = \sum_{\kappa,\lambda=1}^{N} p^i_{\kappa\lambda} x_\kappa^{(n)} x_\lambda^{(n)} \qquad (i = 1, 2, \cdots N). \tag{1}$$

Thus if the distribution $x_i^{(0)}$ is given for an initial generation we can deduce successively the $x_i^{(1)}, x_i^{(2)}, \cdots$ for subsequent generations. Besides, one may wish to express the $x_i^{(n)}$, for any n, explicitly in terms of the initial distribution $x_i^{(0)}$, i.e. to "solve" the system (1). A further problem consists in determining the limit-distribution of the genotypes $\lim_{n\to\infty} x_i^{(n)}$ $(i = 1, \cdots, N)$.

Mendel's heredity theory is based on some ingenious assumptions which are known as Mendel's first and second law. They enable us to define the possible genotypes and to establish the recurrence formula (1); they will be explained and formulated in sections 2 and 3. It is well known that in Mendel's theory it makes an essential difference whether one or more "Mendelian characters" are considered. In the first case Mendel's first law only is used; there are with respect to this character but $N = 3$ different types and the recurrence formula (1) can be derived without difficulty. As early as 1908 G. H. Hardy [5] established the simple but most remarkable result that under random breeding a state of equilibrium is reached in the first filial generation, i.e. $x_i^{(1)} \neq x_i^{(0)}$ (in general) but $x_i^{(n)} = x_i^{(1)}$ $(n = 2, 3, \cdots,)$.[1]

In the case of $m \geqq 2$ Mendelian characters Mendel assumed *independent assortment* of these characters (Mendel's second law). However, within four years after the dramatic rediscovery of Mendel's fundamental paper [10], observations were reported that did not show the results expected for two independent characters. T. H. Morgan [11] and collaborators in basic contributions, con-

[1] See also [12] where the stability of the particular ratio 1:2:1 is recognized.

cluded that a certain *linkage* of genes was to be assumed.[2] Taking that as the starting point, the main purpose of this paper is to establish the basic recurrence formula for the general case of linkage, to solve the corresponding system of difference equations, and to determine the limit distribution of genotypes. Throughout the paper "multiple alleles" are considered instead of making the frequent restriction to two alleles. This generalization is, however, an obvious one (see section 1).

In order to deal with the general problem a *linkage distribution* (l.d.), is introduced. This concept, which seems to be basic to the whole problem, refers to the probability theory of arbitrarily linked events [3]. The *crossover probabilities*, (c.p.), defined by Morgan and Haldane are, notwithstanding their high importance, not sufficient for our purpose. (They turn out to be certain marginal distributions of the l.d.) If, however, $m = 2$ and $N = 10$ (for two alleles), a case studied by W. Weinberg [16] H. S. Jennings [7] and R. B. Robbins [14], the c.p. is equivalent to the l.d. But for the general case the l.d. is needed and the desired results must be derived by other methods than explicit computation, which is feasible if m equals one or two. The original problem of independent assortment appears, of course, as a particular case of the general linkage. This problem was completely solved in 1923 by H. Tietze [15] in a very interesting but rather involved paper. The proof of the limit theorem given in the following pages for the general case is far simpler and shorter than the treatment of the particular case in the older paper and is therefore a simpler proof of Tietze's theorem.

After a brief consideration of the classical case $m = 1$ (section 2), the problem of m arbitrarily linked characters is discussed in section 3 with a particular view to a clear statement of the biological and mathematical assumptions. The l.d., its relation to the c.p., and some basic properties of both are considered in section 4. Then, after a very concise consideration of the case $m = 2$ (section 5), the basic recurrence formula is established in section 6 from which we deduce in section 7 two general limit theorems. The main point is that the limit distribution of genotypes is "uncorrelated" and equals the product of certain marginal distributions of first order deduced from the distribution for the first filial generation. As a kind of an appendix section 8 contains the solution of the system of equations furnished by the general recurrence formula.

In the second part of the paper an attempt is made to contribute to the *linear theory* or theory of the linear order of the genes, from the point of view of probability theory. Accordingly, the linear theory consists in certain assumptions on the l.d., or on an equivalent distribution which will be called *crossover distribution*, (c.d.), and which is more appropriate for this purpose. (Sections 9

[2] "To T. H. Morgan and his associates and students is due the credit for opening up this new field of genetic research; and the small vinegar fly Drosophyla Melanogaster upon which most of their work has been based, has now assumed as great an importance in genetics as the famous peas studied by Mendel." (Sinnot and Dunn, *Principles of Genetics*, New York 1939.)

and 10.) In this connection in section 10 a probability definition of the "distance" d_{ij} of two genes is proposed which, far from being contradictory to Morgan's ideas on the subject seems to formulate them mathematically; (the distance d_{ij} between two genes i and j is defined as the mathematical expectation of the number of crossovers between i and j). This distance is of course additive as it ought to be in the framework of the linear theory.

A problem frequently discussed is whether the crossover probabilities are independent of each other (this independence is not identical with Mendel's free assortment). Observations (see [4a]) did not seem to substantiate this as a general assumption. Then it was concluded that there exists a so-called *interference* which prevents, i.e. diminishes the probability of crossovers too "near" to each other. (See also [13].) It seems to the author that observations on interference should be interpreted in terms of appropriate assumptions regarding the l.d. or the c.d. Again the remark holds that the c.p.'s are not sufficient for describing the situation. Hence in section 11 an attempt is made to understand "interference" by means of the c.d., accepting however the linear theory. It is well known that the explicit presentation of consistent dependent distributions is not trivial (see e.g. [2]). Not many different types of "contagious" distributions are known. In section 11 two such schemes are proposed which, though simple enough, seem to correspond to the general idea of interference. They contain as particular cases the case of independent and the case of disjoint crossovers.

2. One Mendelian character. Hardy's theorem. It will be helpful to start with the simple and well known case of one character introducing the basic concepts in a way appropriate for generalization.

Mendel recognized that the distribution of certain hereditary attributes in organisms is similar to the distribution of attributes in a probability distribution. With respect to a *Mendelian character* each individual is characterized by *two* elements called *genes* which represent two possible alternatives. The color of the flower of peas is such a character, the alternatives being red and white. With respect to this character each plant belongs to one of the three types: red-red, red-white, white-white.[3] These are three different *genotypes*.—In this paper genotypes only will be considered. The difference between genotypes and phenotypes and the related concepts of dominant and recessive qualities will not be dealt with. This is an example of a *two-valued* Mendelian character, i.e. a character for which only two possibilities exist or, using a more technical term,

[3] It will be assumed throughout that the individuals considered are "diploid". That means in the terminology of the preceding example that the only possible types are RR, RW, and WW; or, using A and a: AA, Aa, and aa. Modern research has however revealed that situations may arise where "tetraploids", "hexaploids", etc. briefly "polyploids" prevail, i.e. types like $A^x a^y$ (with $x + y = 2p$). In this case the reproduction cell segregates $A^{x_1} a^{y_1}$ (with $x_1 + y_1 = p$). Stability is no longer reached in the first filial generation. See [4b].

with two alleles. The case of two alleles is most frequently considered in the biological literature where the two possibilities correspond mostly to a dominant and a recessive quality. There is, however, no difficulty in considering from the very beginning the general case of *multiple alleles* where the character under consideration is assumed to be r-valued, i.e. susceptible to r different manifestations (e.g. $r = 5$ possible colors of a plant). These r possible values may be distinguished by the r arguments, $1, 2, \cdots r$.[4]

In the consideration of only one Mendelian character *Mendel's first law* only is used which may be stated as follows:

(a). With respect to one r-valued Mendelian character each individual belongs to one of the $r(r + 1)/2$ possible types, each type being determined by a pair of elements (genes) x and y $\begin{pmatrix} x = 1, \cdots, r \\ y = 1, \cdots, r \end{pmatrix}$.

(b). In the formation of a new individual each parent transmits one of its two genes to the new individual, the other gene coming from the other parent.

(c). The probability for the transmission of either gene is the same and thus equals $\frac{1}{2}$.

We wish to deduce the distribution of genotypes in the $(n + 1)$st generation from the distribution of genotypes in the nth generation *under the assumption of complete panmixia* (random breeding). Moreover, assume that the given initial distributions of genotypes as well as the laws of heredity are the *same for males and females*.[5] In computing successively the new distribution from the preceding one we shall always assume that the distribution of individuals participating in the process of procreation is the same as their distribution when born.

Let us denote a genotype by $(x; y)$, $(x = 1, \cdots, r; y = 1, \cdots, r)$. To fix the ideas we shall assume through this paper that the gene x before the semicolon was transmitted by the mother, and the y after the semicolon by the father of the individual. In some cases which will be considered later this distinction will be relevant. Denote by $w^{(n)}$ $(x; y)$ the probability of the type $(x; y)$ in the nth generation. Since the laws of heredity are the same for males and females we have $w^{(n)}$ $(x; y) = w^{(n)}$ $(y; x)$ and thus have for each generation a symmetric distribution of genotypes with r^m probabilities whose sum is one. There is, however, according to principle (a) no difference between the types $(x; y)$ and $(y; x)$ and therefore it is preferable to group together these types, thus introducing for $x = 1, \cdots, r; y = 1, \cdots, r$:

$$\begin{aligned} v^{(n)}(x; x) &= w^{(n)}(x; x) \\ v^{(n)}(x; y) &= w^{(n)}(x; y) + w^{(n)}(y; x) \text{ where } x < y. \end{aligned} \tag{2}$$

[4] "It is simplest to deal with mere pairs of alternative conditions (alleles) but a theory remains seriously inadequate unless capable of extension to multiple alleles." ([17] p. 224).

[5] It is obvious that we may admit without any change of result different distributions for males and females in the initial generation, as long as random mating takes place afterwards.

Consequently there are $r(r+1)/2$ such probabilities:

$$(2') \quad v^{(n)}(1;1),\ v^{(n)}(1;2),\ \cdots\ v^{(n)}(1;r),\ v^{(n)}(2;2),\ \cdots\ v^{(n)}(2;r),\ \cdots\ v^{(n)}(r;r)$$

where

$$(3) \qquad \sum_{x \leqq y} v^{(n)}(x;y) = 1 \qquad (n = 0, 1, 2, \cdots).$$

Now define $p^{(n)}(x)$ as *the probability that in the nth generation a male (or a female) individual transmits the gene x.* Obviously we have:

$$(4) \qquad \begin{aligned} p^{(n)}(x) = \tfrac{1}{2}\, v^{(n)}(1;x) + \tfrac{1}{2}\, v^{(n)}(2;x) + \cdots + v^{(n)}(x;x) \\ + \tfrac{1}{2}\, v^{(n)}(x;x+1) + \cdots + \tfrac{1}{2}\, v^{(n)}(x;r) \end{aligned}$$

and

$$(4') \qquad \sum_{x=1}^{r} p^{(n)}(x) = 1$$

In fact, the gene x will be transmitted, if an individual possesses this gene and also transmits it. The individuals of type $(y;x)$ (or $(x;y)$) all possess the gene x and transmit it with probability $\frac{1}{2}$ if $y \neq x$ and with probability 1 if $y = x$. Besides, the probability of the type $(x;y)$ in the $(n+1)$st generation is obviously $p^{(n)}(x)p^{(n)}(y)$:

$$(5) \qquad w^{(n+1)}(x;y) = p^{(n)}(x)p^{(n)}(y) = w^{(n+1)}(y;x)$$

or in terms of the $v^{(n)}(x;y)$

$$(5') \qquad \begin{aligned} v^{(n+1)}(x;x) &= [p^{(n)}(x)]^2 \\ v^{(n+1)}(x;y) &= 2p^{(n)}(x)p^{(n)}(y) \qquad (x \leqq y). \end{aligned}$$

Hence by (4) and (5′), $v^{(n+1)}$ has been expressed in terms of $v^{(n)}$ and the recurrence-problem is solved. The distribution $w^{(n+1)}(x;y)(n \geqq 0)$ shows "independence," and is therefore known to be stable. In fact, computing in the same way $p^{(n+1)}(x)$ we get

$$\begin{aligned} p^{(n+1)}(x) &= \tfrac{1}{2}\cdot 2p^{(n)}(1)p^{(n)}(x) + \cdots\ p^{(n)}(x)p^{(n)}(x) \\ &\qquad + \tfrac{1}{2}\cdot 2p^{(n)}(x)p^{(n)}(x+1) + \cdots + \tfrac{1}{2}\cdot 2p^{(n)}(x)p^{(n)}(r) \\ &= p^{(n)}(x)\cdot \sum_{\rho=1}^{r} p^{(n)}(\rho) = p^{(n)}(x) \end{aligned}$$

or

$$(6) \qquad p^{(n+1)}(x) = p^{(n)}(x) \qquad (n = 0, 1, 2, \cdots), (x = 1, 2, \cdots r).$$

This last formula contains G. H. Hardy's famous result [5] *that $p^{(n)}(x)$ is the same for all n:*

$$(7) \qquad p^{(n)}(x) = p^{(0)}(x) \qquad (n = 1, 2, \cdots,)$$

and because of (5′):

$$v^{(n)}(x; y) = v^{(1)}(x; y) \qquad (x \leqq y, n = 2, 3, \cdots). \tag{7′}$$

In case of only one Mendelian property the distribution of genotypes reaches a stationary state in the first filial generation.

3. Basic assumptions in case of m Mendelian characters. A new situation presents itself if there is more than one character. In case of m characters a genotype is described by $2m$ numbers $(x_1, \cdots, x_m; y_1, \cdots, y_m)$ or briefly $(x; y)$ (e.g. for $m = 5$, $r = 9$: $(1,2,3,4,6; 2,7,3,5,9)$. There are primarily $N = r^{2m}$ possible types because on each of the $2m$ places any of the r numbers can be written. Now, if the types $(x; y)$ and $(y; x)$ are considered as identical genotypes, the number of different genotypes reduces to $N_1 = \frac{r^m}{2}(r^m + 1)$ (e.g. for $r = 2$, $m = 1$: $N_1 = 3$; for $r = m = 2$: $N_1 = 10$). It is essential for the understanding of linkage that in counting this way two types like $(1,3; 5,7)$ and $(1,7; 5,3)$ or $(1,1;2,2)$ and $(1,2;2,1)$ are considered as different although in both cases the individual possesses with respect to the first character the gene pair $1,5$ and with respect to the second the pair $3,7$. If no difference is assumed between two such types the number of different genotypes reduces to $N_2 = \left(\frac{r(r+1)}{2}\right)^m$. (E.g. for $r = 2$: $N = 4^m$, $N_1 = \frac{1}{2}\cdot 2^m(2^m + 1)$, $N_2 = 3^m$; hence for $m = r = 2$: $N = 16$, $N_1 = 10$, $N_2 = 9$ or for $r = 2$, $m = 3$: $N = 64$, $N_1 = 36$, $N_2 = 27$). Which method of counting is the correct one?

The answer is that there are but N_2 different genotypes if *Mendel's second law*, the *law of independent assortment*, is accepted. Then and only then there is no difference between types like $(1,3:5,7)$ and $(1,7:5,3)$. Under the assumption of general linkage however, these types must be distinguished, not as individuals, but with respect to their heredity properties, i.e. considered as parents of a new generation. Under this assumption there are in general N_1 different types. This will be discussed presently in more detail.

Let us first consider Mendel's original theory as contained in his *first and second law*. Analogous to (a), (b) and (c) in §2 we now formulate as follows:

(a′) With respect to m characters the genotype of an individual is characterized by m pairs of numbers. Two individuals are of the same type if to each of the m characters corresponds the same pair. Hence there are $N_2 = \left(\frac{r(r+1)}{2}\right)^m$ genotypes.

(b′) In the formation of a new individual a parent of type $(x_1, \cdots, x_m; y_1, \cdots, y_m)$ transmits to the offspring, corresponding to each of the m characters, one of the two genes which he (or she) possesses with respect to this character.

(c′) The probability of transmitting any of these 2^m combinations is the same and therefore equal to $1/2^m$.

Consider e.g. the individual (1,2,3:1,4,7); the pair 1,1 corresponds to the first character the pair 2,4 to the second and 3,7 to the third. Under the assumptions of Mendel's original theory this individual is of the same type with (1,4,3; 1,2,7) and (1,2,7; 1,4,3), and of course with (1,4,7; 1,2,3), etc. As $m = 3$, it may transmit eight combinations which in the preceding example reduce to four, because the individual is homozygous in the first character. These four combinations are 1,2,3 or 1,4,3 or 1,2,7 or 1,4,7 each with probability $2 \times \frac{1}{8} = \frac{1}{4}$.

The distribution of genotypes in successive generations under the assumption of Mendel's second law has been investigated by H. Tietze [15] who also considers the limiting distribution as $n \rightarrow \infty$. His results will appear as a particular case of our general considerations.

In order to discuss the basic facts which lead to the idea of linkage let us for the moment consider the case $m = 2$. Soon after the rediscovery of Mendel's work Bateson and Punett reported observations which did not give the expected numerical results. To understand the type of such an observation assume that a homozygous male of type (1,1; 1,1), [or any other homozygous type, e.g. (2,3; 2,3)] is mated to a homozygous female of type (2,2; 2,2) [or to any homozygous type different from the first e.g. (4,5; 4,5)]. Obviously, in this case there is only one possible kind of offspring namely (2,2; 1,1), [or (4,5; 2,3)]. But if now one of these daughters is mated to a homozygous male of the original type (1,1; 1,1), there are four kinds of possible offspring, namely (2,2; 1,1), (2,1; 1,1), (1,2; 1,1), and (1,1; 1,1), corresponding to the four combinations of genes transmitted by the heterozygous (dihybrid) daughter [or (4,5; 2,3), (4,3; 2,3), (2,5; 2,3), and (2,3; 2,3)]; and according to the idea of free assortment each of these four combinations should appear with the same relative frequency: $\frac{1}{4}$. But it was observed that the combined frequency of the two types (1,1; 1,1) and (2,2; 1,1) was larger than that of the types (2,1; 1,1) and (1,2; 1,1). "The characters that went in together have come out together in a much higher percentage than expected from Mendel's second law, viz. the law of independent assortment" [11]. Morgan, in his theory of the gene called this "tendency" *linkage*. The idea is that the two genes 2,2 and 1,1 which have been together in the maternal individual tend to stay together and that nature has to make an effort to produce a so-called *crossing-over*, i.e. a separation of the genes "that came in together,"—such that a female of type (2,2; 1,1) may transmit the group 1,2 or the group 2,1. In other words, *the idea of linkage implies an influence of the grandparents.*

According to observation the percentage of crossing over varies from 0 to 50 per cent, i.e. from *complete linkage* to *free assortment*. It will appear however that in principle crossover-values greater than 50 per cent cannot be excluded. It was also observed that the percentage of individuals of type (1,1; 1,1) equals very nearly that of individuals of type (2,2; 1,1), as we would expect. In the same way the percentages of types (2,1; 1,1) and (1,2; 1,1) are nearly equal, their sum yielding the *crossover-ratio*. Hence the four probabilities corresponding to the formation of the four types (1,1; 1,1), (2,2; 1,1), (2,1; 1,1), and (1,2; 1,1) are assumed to be $(1 - c)/2$, $(1 - c)/2$, $c/2$, and $c/2$. It is important to notice that these are at the same time the probabilities that the female of type (2,2; 1,1) (which was mated to the homozygous (1,1; 1,1), transmits the groups 1,1 or 2,2 or 2,1 or 1,2 respectively.

In the general case of m characters there are $\binom{m}{2} = \frac{m(m-1)}{2}$ crossover probabilities. In this case Morgan assumes *linkage-groups*, each group consisting of m_i elements with $\sum_i m_i = m$, such that "there is linkage between the elements of each group but that the members belonging to different linkage groups assort independently, in accordance with Mendel's second law." This idea will be reconsidered in Section 4.

If we now wish to solve our first basic problem, i.e. to derive the distribution of genotypes in any later generation from an initial distribution of genotypes, then the concept of crossover probabilities does not suffice. The complex possibilities which arise if Mendel's second law is no longer accepted as universally valid cannot be adequately described in terms of crossover probabilities. Or, more exactly: It will be seen that if $m \geqq 4$ the crossover probabilities are no longer sufficient, whereas for $m = 2$ and $m = 3$ this concept is general enough. For the complete description of the hereditary mechanism in the general case a so-called *linkage distribution*, l.d., is needed which involves 2^m probabilities with sum equal to one. Let us define this distribution.

Consider an individual of type $(x_1, \cdots, x_m; y_1, \cdots, y_m) \equiv (x; y)$, where the x are the maternal genes, the genes contributed by the mother of the individual, and the y the paternal genes. Denote by S the set of the m numbers $1, 2, \cdots, m$, by A any subset of S, and by A' the complementary subset $A' = S - A$. *Denote by $l(A)$ the probability that the individual $(x; y)$ transmits the paternal genes belonging to A and the maternal genes belonging to A'.* There are $1 + m + \binom{m}{2} + \cdots + 1 = 2^m$ such subsets A and accordingly 2^m probabilities $l(A)$ where

$$\sum_{(A)} l(A) = 1. \tag{8}$$

In accordance with the previously reported observations and with our assumption of equal conditions for both sexes one must assume that

$$l(A) = l(A'). \tag{8'}$$

The conditions (8) and (8′) reduce the number of freely disposable values of the l-distribution to $(2^{m-1} - 1)$. The l-distribution is a socalled *m-dimensional* or *m-variate* alternative which could also and occasionally will be denoted by $l(\epsilon_1, \epsilon_2, \cdots, \epsilon_m)$ where $\epsilon_i = 0$ or 1. Thus e.g. $l(1,1,1,0,0,1)$ is the probability that the genes $y_1, y_2, y_3, x_4, x_5, y_6$, are the genes contained in the germ cell of the individual $(x; y)$. Here the set A consists of the numbers 1,2,3,6, and A' of 4,5.

Analogous to the statements (a), (b), (c) of §2 and (a′), (b′), (c′) of the present section we may now formulate the *principles of Mendel's theory of heredity under the assumption of a possible linkage of the genes:*

(a″) With respect to m Mendelian characters an individual is characterized by two sets of numbers each consisting of m numbers, viz. $x_1, \cdots, x_m$ and $y_1, \cdots, y_m$, where $\begin{matrix} x_i \\ y_i \end{matrix} = 1, 2, \cdots, r$. If the type of an individual is designated by $(x_1, \cdots, x_m; y_1, \cdots, y_m) \equiv (x; y)$ where x and y denote the maternal and paternal contributions respectively, then $(x; y) = (y; x)$. Hence there are $N_1 = \frac{1}{2}r^m(r^m + 1)$ types of individuals.

(b″) ≡ (b′) In the formation of a new individual each parent transmits to the offspring one set of m genes.

(c″) For each parent, the 2^m probabilities of transmitting any one of these 2^m possible sets are given by a linkage distribution $l(A)$ where A is a subset of the set S consisting of the m numbers $1, 2, \cdots, m$, and $l(A)$ is the probability that the transmitted set consists of the paternal genes belonging to A and of the maternal genes belonging to $A' = S - A$, and $l(A) = l(A')$.

4. Some properties of the linkage distribution and of the crossover probabilities. In the following we shall need marginal distributions, that is partial sums, of the probabilities within a distribution. In a usual notation:

$$(9)\quad \begin{aligned} l_1(x_1) &= \sum_{x_2} \sum_{x_3} \cdots \sum_{x_m} l(x_1, x_2, \cdots, x_m), \ldots\ldots\ldots\ldots\ldots\ldots m \text{ distributions} \\ l_{12}(x_1, x_2) &= \sum_{x_3} \cdots \sum_{x_m} l(x_1, x_2, \cdots, x_m), \ldots\ldots\ldots\ldots \binom{m}{2} \text{ distributions} \\ l_{123\cdots m-1}(x_1, x_2, \cdots, x_{m-1}) &= \sum_{x_m} l(x_1, x_2, \cdots, x_m), \ldots\ldots\ldots m \text{ distributions} \\ l_{12\cdots,m}(x_1, x_2, \cdots, x_m) &= l(x_1, x_2, \cdots, x_m) \ldots\ldots \text{the original distribution.} \end{aligned}$$

These are general formulae for any discontinuous distribution. But if the distribution happens to be an alternative, as the l.d., where x_i takes only two values, any marginal distribution can be completely characterized by two subsets A and A_1 of S where $A \supset A_1$. *Denote by $l_A(A_1)$ the sum of all possible linkage probabilities which contain all points of A_1 and no point of $A - A_1$.* If, e.g. $m = 8$ and A consists of 1,3,5,6 and A_1 of 1,3,6 then $l_A(A_1) = l_{1356}\,(1,1,0,1) = \sum_{x_2, x_4, x_7, x_8} l(1, x_2, 1, x_4, 1, 0, x_7, x_8)$. According to the previous notation we have as usual

$$(10)\qquad l_S(A_1) = l(A_1), \text{ or } l_{1,2,\ldots,m}(x_1, \cdots, x_m) = l(x_1, \cdots, x_m)$$

and

$$l_O(O) = 1, \text{ if } A = O \text{ is empty.}$$

We will use for the linkage distribution and their marginal distributions the customary notations or these new notations, whichever is more convenient.[6]

As an immediate consequence of our definitions we get the following properties of the l.d.

(i) *If* (8) *holds for any A than*

$$l_A(A_1) = l_A(A - A_1). \tag{11}$$

(ii) *As a consequence of* (8) *it follows* (*with the notation* (9)) *that*

$$l_i(1) = l_i(0) = \tfrac{1}{2}. \tag{9'}$$

(iii) If c_{ij} denotes the c.p. between i and j, then

$$c_{ij} = c_{ji} = l_{ij}(1,0) + l_{ij}(0,1) = 2l_{ij}(1,0) = 2l_{ij}(0,1). \tag{12}$$

(iv) *For any three subscripts i, j, k the "triangular" relation holds*

$$c_{ij} + c_{jk} \geqq c_{ik} \tag{13}$$

and

$$c_{ij} + c_{jk} + c_{ik} \leqq 2. \tag{14}$$

To prove this consider the marginal distribution $l_{ijk}(x_i x_j x_k)$. From (11) and (12) we conclude

$$\begin{aligned} c_{ij} &= 2[l_{ijk}(100) + l_{ijk}(010)] \\ c_{ik} &= 2[l_{ijk}(100) + l_{ijk}(001)] \\ c_{jk} &= 2[l_{ijk}(010) + l_{ijk}(001)] \\ 1 &= 2[l_{ijk}(000) + l_{ijk}(100) + l_{ijk}(010) + l_{ijk}(001)]. \end{aligned}$$

[6] It is easy to indicate experiments which should furnish the relative frequencies corresponding to the l.d.: If a homozygous female $(x_1, \cdots, x_m; x_1, \cdots, x_m)$ is mated to a homozygous male $(y_1, \cdots, y_m; y_1, \cdots, y_m)$ where each $x_i \neq y_i$, the resulting offsprings will all be of type $(x_1, \cdots, x_m; y_1, \cdots, y_m)$. If such an offspring is back crossed to $(y_1, \cdots, y_m; y_1, \cdots, y_m)$ there will be 2^m different genotypes of offsprings, viz. $(x_1, x_2, \cdots, x_m; y_1, y_2, \cdots, y_m)$, $(y_1, x_2, \cdots, x_m; y_1, y_2, \cdots, y_m)$, etc. whose frequencies are proportional to the 2^m values of the l.d., viz. to $l(0, 0, \cdots, 0)$, $l(1, 0, 0, \cdots, 0)$ etc. Such an experiment should give the same results for any two sets of x's and y's. (There is, of course, the statistical problem how to determine the "best" values of the l.p. from these observations.) In an analogous way a marginal distribution can be observed: Suppose we wish for $m = 5$, the $l_{123}(\epsilon_1, \epsilon_2, \epsilon_3)$. The offspring of a cross between females $(x_1, x_2, x_3, x_4, x_5; x_1, x_2, x_3, x_4, x_5)$ and males $(y_1, y_2, y_3, x_4, x_5; y_1, y_2, y_3, x_4, x_5)$ are of type $(x_1, x_2, x_3, x_4, x_5; y_1, y_2, y_3, x_4, x_5)$. If they are crossed to $(x_1, \cdots, x_5; x_1, \cdots, x_5)$ there will be eight different types of offsprings proportional to the eight values of $l_{123}(\epsilon_1, \epsilon_2, \epsilon_3)$. In this last setup the y_i should be dominant and in the experiment, described above, the y_i should be recessive in order to be able to distinguish between the phenotypes of the individuals.

Solving these equations with respect to the l-values we get

$$\begin{aligned} l_{ijk}(100) &= \tfrac{1}{4}\,(c_{ij} + c_{ik} - c_{jk}) \\ (15) \qquad l_{ijk}(010) &= \tfrac{1}{4}\,(c_{ij} + c_{jk} - c_{ik}) \\ l_{ijk}(001) &= \tfrac{1}{4}\,(c_{ik} + c_{jk} - c_{ij}) \\ (16) \qquad l_{ijk}(000) &= \tfrac{1}{4}\,(2 - c_{ij} - c_{ik} - c_{jk}). \end{aligned}$$

Thence (13) and (14) follow. The condition (14) is of course always fulfilled if $c_{ij} \leqq \frac{1}{2}$, but this restriction does not seem to be necessary. From (15) and (16) we deduce:

(v) *If* $m = 3$, *the set of three c.p.* c_{12}, c_{13}, c_{23} *for which the inequalities* (13), (14) *hold is equivalent to the l.d.* $l(x_1, x_2, x_3)$ *for which* (8) *holds.* For $m \geqq 4$ the c.p. are no longer equivalent to the l.d. Another necessary condition for the c.p. will be derived in section 8.

Now let us consider and characterize *some important particular cases of the l.d.*

(i) *Free assortment* (Mendel). In this case all 2^m values of the l.d are equal and therefore equal to $(\frac{1}{2})^m$.

(ii) *Complete linkage* (reported by Morgan and other authors). In terms of the l.d. this means

$$(17) \qquad l(1,1, \cdots, 1,1) = l(0,0, \cdots, 0,0) = \tfrac{1}{2} \text{ or } l_S(S) = \tfrac{1}{2}.$$

Consequently, all other values of the l.d. are zero. It follows that *all c.p. are zero* because all $l_{ij}\,(1,0)$ are zero. (See also Theorem I, section 7.)

(iii) *Linkage groups* (Morgan). In terms of the l.d. this means that *the l.d. resolves into a product of several distributions*, e.g.

$$(18) \qquad l(x_1, x_2, \cdots, x_9) = f(x_1, x_2)g(x_3, x_4, x_5)h(x_6, x_7, x_8, x_9).$$

(There is no loss of generality in assuming that numerically consecutive characters form a linkage group.) As f, g, and h are distributions it follows with notation (9) that "within" the groups:

$$c_{12} = 2f(10), \qquad c_{34} = 2g_{34}(10), \cdots, \qquad c_{45} = 2g_{45}(10),$$
$$c_{67} = 2h_{67}(10), \cdots, \qquad c_{89} = 2h_{89}(10)$$

these crossover values are quite arbitrary. On the other hand we have because of (9′)

$$f_i(1) = f_i(0) = g_j(1) = g_j(0) = h_k(1) = h_k(0) = \tfrac{1}{2},$$
$$(i = 1, 2;\ j = 1, 2, 3;\ k = 1, \cdots 4)$$

Hence for the c.p. "among" the groups

$$c_{13} = 2\cdot\tfrac{1}{2}\cdot\tfrac{1}{2} = \tfrac{1}{2}, \text{ etc. Hence } c_{13} = c_{14} = c_{23} = \cdots = c_{59} = \tfrac{1}{2}$$

in exact accordance with Morgan's idea of linkage groups. If each group consists of only one element: $l(x_1, x_2, \cdots, x_m) = f(x_1)g(x_2) \cdots k(x_m)$ it follows that $f(x_1) = g(x_2) = \cdots = k(x_m) = \frac{1}{2}$ for $x_i = 0, 1$, hence $l(x_1, \cdots, x_m) = (\frac{1}{2})^m$ for all combinations of the arguments and we have again free assortment.

(*iv*) *Groups of completely linked characters.* Combining and generalizing the ideas of (ii) and (iii) we may speak of i groups of completely linked characters if *within such a group no crossover takes place.* Then the m_i characters in each group act as one character. An example will suffice. Suppose $m = 9$ and three such groups, consisting of the characters 1,2, and 3,4,5, and 6,7,8,9 respectively. Assume that

$$l(11, 111, 1111) = l(00, 000, 0000) = a, \quad l(00, 111, 1111) = l(11, 000, 0000) = C_1$$
$$l(11, 000, 1111) = l(00, 111, 0000) = C_2, \quad l(11, 111, 0000) = l(00, 00, 1111) = C_3$$

where these four numbers are $\neq 0$ and with sum $\frac{1}{2}$; hence all other probabilities are zero. It follows that the c.p. "within" the groups are all zero: $c_{12} = c_{34} = \cdots = c_{45} = c_{67} = \cdots = c_{89} = 0$, but the "among" c.p. are different from zero, e.g. $c_{13} = c_{14} = c_{15} = c_{23} = c_{24} = c_{25} = 2C_1 + 2C_2$ and, with an obvious notation: $c_{I,II} = 2(C_1 + C_2)$, $c_{I,III} = 2(C_1 + C_3)$, $c_{II,III} = 2(C_2 + C_3)$.

A particular case (also a particular case of (iii)) arises if the l.d. resolves into a product of some distributions such that there is complete linkage in each of these. The "within" crossovers are then again zero but all the c.p. "among" the groups equal $\frac{1}{2}$.

5. The case $m = 2$. It will be easier for the reader if this case, though it has been investigated before by several authors [16], [7], [14], will be presented by means of explicit computations before attempting the general one where m and r are arbitrary.

If $m = r = 2$, the number of types $(x_1, x_2; y_1, y_2)$ equals ten. The l.d. is completely determined by the c.p. $c_{12} = c$ and v.v., because $l(10) = l(01) = c/2$, $l(00) = l(11) = (1 - c)/2$. Now let $p^{(n)}(x_1, x_2)$ be the probability that in the nth generation a male (or female) individual *transmits the genes* x_1, x_2; and denote by $p_1^{(n)}(x_1)$ and $p_2^{(n)}(x_2)$ the respective marginal distributions. The formula corresponding to (4) then becomes

$$(19)\quad \begin{aligned} p^{(n)}(1,1) &= v^{(n)}(1,1;1,1) + \tfrac{1}{2}v^{(n)}(1,1;1,2) + \tfrac{1}{2}v^{(n)}(1,1;2,1) \\ &\quad + \frac{1-c}{2}v^{(n)}(1,1;2,2) + \frac{c}{2}v^{(n)}(1,2;2,1), \end{aligned}$$

and three analogous formulae. To understand this, consider e.g. the last term of (19); it is the probability that an individual be of type (1,2; 2,1) or (2,1; 1,2) and transmits the set (1,1). By (19) $p^{(n)}(x_1, x_2)$ is deduced from the given distribution $v^{(n)}$ of genotypes.

If, as before, x and y are written for x_1, x_2 and y_1, y_2 it is to be understood that $x = y$ means $x_1 = y_1$ and $x_2 = y_2$. The relation corresponding to (5′) takes then the form

$$(20)\quad \begin{aligned} v^{(n+1)}(x; y) &= p^{(n)}(x)p^{(n)}(x) \quad \text{if} \quad x = y \\ &= 2p^{(n)}(x)p^{(n)}(y) \quad \text{if} \quad x \neq y. \end{aligned}$$

Applying (19) to the $(n + 1)$st generation and using (20) we get the recurrence formula

$$(21)\quad \begin{aligned} p^{(n+1)}(1,1) = [p^{(n)}(1,1)]^2 + p^{(n)}(1,1)p^{(n)}(1,2) + p^{(n)}(1,1)p^{(n)}(2,1) \\ + (1 - c)p^{(n)}(1,1)p^{(n)}(2,2) + cp^{(n)}(1,2)p^{(n)}(2,1). \end{aligned}$$

Here the right side can be rewritten so as to give

$$(22)\quad p^{(n+1)}(1,1) = (1 - c)p^{(n)}(1,1) + cp_1^{(n)}(1)p_2^{(n)}(1)$$

and three analogous formulae. Because of (7):

$$(22')\quad p^{(n+1)}(x_1, x_2) = (1 - c)p^{(n)}(x_1, x_2) + cp_1^{(0)}(x_1)p_2^{(0)}(x_2).$$

From this recurrence formula, which has the particularly simple property that the second term on the right side is independent of n, it is easy to derive step by step:

$$(23)\quad p^{(n)}(x_1, x_2) = (1 - c)^n p^{(0)}(x_1, x_2) + [1 - (1 - c)^n]p_1^{(0)}(x_1)p_2^{(0)}(x_2).$$

Hence, if $c \neq 0$:

$$(24)\quad \lim_{n\to\infty} p^{(n)}(x_1, x_2) = p_1^{(0)}(x_1)p_2^{(0)}(x_2).$$

The preceding results were obtained by Robbins and Jennings. We will formulate a theorem after having studied the general case of arbitrary m and r.[7]

6. The general recurrence formula. Considering random mating and assuming general linkage, we now wish to find the relations which correspond to the formulae (19)–(22) in the case of m r-valued characters. It will turn out, that, by using the l.d., the following proof of the general case becomes surprisingly simple compared with older investigations of the particular case of free assortment, the values of the l.d. acting somehow as natural "separators" for certain groups of terms.

Denote by $w^{(n)}(x_1, \cdots, x_m; y_1, \cdots, y_m) \equiv w^{(n)}(x; y)$ the probability of a genotype whose maternal genes are the x and whose paternal genes the y. Then from (a''):

$$(25)\quad w^{(n)}(x;\ y) = w^{(n)}(y;\ x).$$

Writing $x = y$ if and only if $x_i = y_i$, $(i = 1, \cdots, m)$ we put just as in (2)

$$(25')\quad \begin{aligned} v^{(n)}(x; y) &= w^{(n)}(x; x), && \text{if } x = y \\ &= w^{(n)}(x; y) + w^{(n)}(y; x) = 2w^{(n)}(x; y), && \text{if } x \neq y. \end{aligned}$$

[7] A suggestive remark, repeatedly made by Professor S. Wright states that (assuming random mating) there can be no equilibrium until all of the factors are combined at random. This is indeed a necessary condition for stability.

There are r^{2m} w-values and $\frac{1}{2}r^m(r^m + 1)$ v-values in each generation the respective sums being always equal to one. Denote by $p^{(n)}(x_1, \cdots, x_m)$ *the probability that a male (female) individual of the nth generation transmits the genes* x, and by

$$p_i^{(n)}(x_i),\ p_{ij}^{(n)}(x_i, x_j),\ \cdots,\ p_{12\cdots m}^{(n)}(x_1, \cdots, x_m) = p^{(n)}(x_1, \cdots, x_m)$$

the corresponding marginal distributions, defined as usual (see (9)). Sometimes it will be convenient to denote such a marginal distribution by $p_A(z_A) \equiv p_A(z)$ where $A \subset S$, and $p_A(z)$ is the sum of all $p(x)$ such that $x_i = z_i$ for all $i \epsilon A$. Following convention the subscript will be omitted if $A = S$; hence $p_S(z) = p(z)$ and if A is empty, $A = 0$, the corresponding $p_0(z) = 1$.

To simplify the writing $p(x)$, $v(x; y)$, etc. will be written instead of $p^{(n)}(x)$, $v^{(n)}(x; y)$, etc. and $p'(x)$, $v'(x; y)$, etc. for $p^{(n+1)}(x)$, etc. Finally, remember that $l(A)$ is the probability that the paternal genes of A and the maternal genes of $A' = S - A$ will be transmitted and accordingly $l_A(A_1)$ is the (marginal) probability that the paternal genes of A_1 and the maternal genes of $A - A_1$ will be transmitted. $(S \supset A \supset A_1)$.

Let us derive $p'(z)$ from $p(z)$. From the meaning of the different distributions we gather that

$$p(z) = \Sigma l(A) w(x; y) \tag{26}$$

where A is an arbitrary subset of S and x and y such that

$$\begin{aligned} y_i &= z_i \quad \text{for} \quad i \epsilon A \\ x_i &= z_i \quad \text{``} \quad i \epsilon A'. \end{aligned} \tag{a}$$

In fact, the set z will be transmitted if and only if an individual possesses these genes and also transmits them; now consider any $l(A)$ i.e. the probability to transmit the paternal genes of A; this probability is to be multiplied by all possible w-probabilities which contain as arguments the paternal genes of A and the maternal genes of A', as stated in (a). Now let us write (26) also for the $(n + 1)$st generation:

$$p'(z) = \Sigma l(A) w'(x; y). \tag{26'}$$

Next we have, just as always, [see (5), (20)]

$$w'(x; y) = w'(y; x) = p(x)p(y). \tag{27}$$

Hence from (26') and (27) follows

$$p'(z) = \Sigma l(A) p(x) p(y) \tag{28}$$

with the condition of summation given by (a).

The right side of (28) contains $(2r)^m$ terms. Now we will write it in two different ways by collecting its terms under two different aspects: (i) arranged according to the marginal values of the l-distribution (ii) arranged according to the marginal values of the p-distribution. Let us begin with (i).

The genes $z_1, z_2, \cdots, z_m$ can be transmitted only by individuals which possess each z_i either before or after the semicolon or both; (either from the mother or from the father or from both parents). Hence, if A_1 and A_2 are two disjoint subsets of S, the type of such an individual is such that

$$\text{(b)} \qquad \begin{array}{lll} x_i \neq z_i, & y_i = z_i & \text{for all} \quad i \in A_1 \\ x_j = z_j, & y_j \neq z_j & \text{`` \quad ``} \quad j \in A_2 \\ x_k = y_k = z_k & & \text{`` \quad ``} \quad k \in S - A_1 - A_2 . \end{array}$$

Hence the paternal genes of A_1 and the maternal genes of A_2 must be transmitted and for the remaining genes either choice is admissible. Consequently, each $w'(x; y)$ in (26′)—or, what is the same, each $p(x)p(y)$ in (28)—is multiplied by the probability that the paternal genes of A_1 and the maternal genes of A_2 are transmitted. Now writing $A_1 + A_2 = A$ this last probability is exactly the marginal probability $l_A(A_1) = l_A(A_2)$. Thence

$$\text{(29)} \qquad p'(z) = \Sigma p(x)p(y)l_A(A_1)$$

where the sum is extended over all pairs x, y defined by (b). This is a first recurrence formula. If in (29) $w'(x; y)$ is written instead of $p(x)p(y)$ and then all accents are omitted we get

$$\text{(30)} \qquad p(z) = \Sigma w(x; y)l_A(A_1)$$

with the summation according to (b). This formula is necessary in order to derive $p(z)$ from the given distribution $w(x; y)$ of genotypes. It corresponds to (19).

Now let us collect the terms of (28) in the second way. Let us determine the factor of any $l(A)$ in (28), e.g. of $l(1,1,0,0,0)$ (where $m = 5$ and A the subset 1,2). Any factor of $l(1,1,0,0,0)$ must be of the form $p(z_1, z_2, \cdot, \cdot, \cdot)$ $p(\cdot, \cdot, z_3, z_4, z_5)$ where all possible values of the variables must be written on the empty places marked by points, and the sum of all these products is to be taken. Now, as in each of the two p's on each of the free places all numbers between 1 and r have to be used, the sum of all these products resolves into the product of the respective sums of the p's. In such a sum each term, on the places belonging to A contains the same fixed values z_A and on the other places any possible value combination; hence such a sum is precisely the marginal probability $p_A(z_A) = p_A(z)$ and the same holds for the other sum of the p's and for $A' = S - A$. Thus we get the second, even more important recurrence formula

$$\text{(31)} \qquad p'(z) = \sum_{(A)} l(A)p_A(z)p_{A'}(z)$$

where the sum is over all subsets A of S. This formula corresponds to (22) and the limit theorem which will be proved in the next section is an almost immediate consequence of (31). It is worth noticing that the derivations of

(31) and (29) from (28) are completely independent of each other and that only (31) is needed for the limit theorem

From (29) and (31) the interesting identity follows

$$\sum_{(A)} l(A)p_A(z)p_{A'}(z) = \Sigma p(x)p(y)l_A(A_1) \tag{32}$$

which somehow reminds us of a general Abel-transformation.

Let us summarize: (i) From a given distribution of genotypes $w^{(n)}$ (or $v^{(n)}$) the $p^{(n)}$ are derived by (30). (ii) From these $p^{(n)}$ the $w^{(n+1)}$ follow by (27) (or $v^{(n+1)}$ by (25′) and (27)). (iii) Instead of step (ii), from $p^{(n)}$ the consecutive $p^{(n+1)}, p^{(n+2)}, \cdots, p^{(n+\nu)}$ may be derived directly by means of (31). Finally, if desired, $w^{(n+\nu+1)}$ follows by (27).

As an illustration of these formulae let us write (31) for $m = 3,4,5$:

$$\begin{aligned} p'(x_1, x_2, x_3) = 2[&l(000)p(x_1, x_2, x_3) + l(100)p_1^{(0)}(x_1)p_{23}(x_2, x_3) \\ &+ l(010)p_2^{(0)}(x_2)p_{13}(x_1, x_3) \\ &+ l(001)p_3^{(0)}(x_1)p_{12}(x_1, x_2)] \end{aligned} \tag{31′}$$

$$\begin{aligned} p'(x_1, x_2, x_3, x_4) = 2[&l(0000)p(x_1, x_2, x_3, x_4) \\ &+ l(1000)p_1^{(0)}(x_1)p_{234}(x_2, x_3, x_4) + \cdots \\ &+ l(1100)p_{12}(x_1, x_2)p_{34}(x_3, x_4) \\ &+ l(1010)p_{13}(x_1, x_3)p_{24}(x_2, x_4) \\ &+ l(1001)p_{14}(x_1, x_4)p_{23}(x_2, x_3)] \end{aligned} \tag{31″}$$

$$\begin{aligned} p'(x_1, x_2, x_3, x_4, x_5) = 2[&l(00000)p(x_1, \cdots, x_5) \\ &+ l(10000)p_1^{(0)}(x_1)p_{2345}(x_2, x_3, x_4, x_5) + \cdots \\ &+ l(11000)p_{12}(x_1, x_2)p_{345}(x_3, x_4, x_5) + \cdots]. \end{aligned} \tag{31‴}$$

In the last formula the last group contains ten terms. As an illustration of (30) we write e.g. for $m = 3$, $r = 2$, with $p^{(n)} = p$ and $v^{(n)} = v$:

$$\begin{aligned} p(x_1, x_2, x_3) \equiv v(x_1, x_2, x_3 ; x_1, x_2, x_3) + \tfrac{1}{2}[&v(x_1, x_2, x_3 ; y_1, x_2, x_3) \\ &+ v(x_1, x_2, x_3 ; x_1, y_2, x_3) + \cdots] \\ &+ [l_{12}(00)v(x_1, x_2, x_3 ; y_1, y_2, x_3) \\ &+ l_{13}(00)v(x_1, x_2, x_3 ; y_1, x_2, y_3) + \cdots] \\ &+ l(000)v(x_1, x_2, x_3 ; y_1, y_2, y_3) \\ &+ [l(100)v(y_1, x_2, x_3 ; x_1, y_2, y_3) \\ &+ l(010)v(x_1, y_2, x_3 ; y_1, x_2, y_3) + \cdots]. \end{aligned} \tag{30′}$$

7. Limit theorems. In order to find $\lim_{n\to\infty} p^{(n)}(x_1, \cdots, x_m)$ we write the recurrence formula (31) in the form

$$(33) \qquad p^{(n+1)}(x) - 2l(00 \cdots 0)p^{(n)}(x) = \sum_{(A)}{}' \, l(A)p_A^{(n)}(z)p_{A'}^{(n)}(z).$$

Here $\sum_{(A)}'$ means a sum over all subsets A of S which are neither void nor equal to S. If we write $q_m^{(n)}$ for the right side of (33) and $p^{(n)}(x) = p_m^{(n)}$, $2l(0, \cdots, 0) = \alpha_m$ the last equation takes the form

$$(34) \qquad p_m^{(n+1)} - \alpha_m p_m^{(n)} = q_m^{(n)}.$$

Consider first the case $\alpha_m = 1$, or $l(0, \cdots, 0) = l(1, \cdots, 1) = \frac{1}{2}$, i.e. *complete linkage*, as defined in section 3. In this case all $l(A)$-values on the right side of (33) are zero, hence $q_m^{(n)} = 0$ and

$$(35) \qquad p_m^{(n+1)} = p_m^{(n)} \qquad (n = 0, 1, 2, \cdots).$$

This is exactly the same result as (7): All $p_m^{(n)}$ are equal to $p_m^{(0)}$ and because of (27) also

$$(36) \qquad w^{(n)}(x; y) = w^{(1)}(x; y) \quad \text{or} \quad v^{(n)}(x; y) = v^{(1)}(x; y) \quad (n = 1, 2, \cdots).$$

In fact, if the characters are completely linked, they act as one character. Hence we have

THEOREM I. *If the m Mendelian characters are completely linked, the distribution of genotypes reaches the stationary state in the first filial generation.*

Now consider (34) in the general case where $0 \leqq \alpha_m < 1$. Then the following *lemma* will be used: *If in a recurrence formula of the form* (34), $|\alpha_m| < 1$ *and* $\lim_{n\to\infty} q_m^{(n)} = q_m$ *exists, then* $\lim_{n\to\infty} p_m^{(n)} = p_m = q_m/(1 - \alpha_m)$. This can be proved directly in various simple ways. It may also be regarded as a consequence of well-known general convergence theorems. See also [15].

In order to apply the lemma let us first notice that q_2 exists. In fact, $p^{(n+1)}(x_1, x_2) - 2l(00)p^{(n)}(x_1, x_2) = 2l(01)p_1^{(0)}(x_1)p_2^{(0)}(x_2)$ and as the right side is independent of n, q_2 certainly exists. Hence, it follows from the lemma that p_2 exists. For $m = 3$ the recurrence formula (31′) shows that $q_3^{(n)}$ contains no marginal distribution of p of an order higher than two; therefore each of the terms of $q_3^{(n)}$ approaches a limit, hence $q_3 = \lim_{n\to\infty} q_3^{(n)}$ exists, and consequently, because of the lemma, p_3 exists. We may continue in this way because in (33) all marginal distributions of p on the right side are of an order $\leqq m - 1$. Hence for every m the $q_m^{(n)}$ approaches a limit and consequently the $\lim_{n\to\infty} p_m^{(n)}$ exists. Finally, in order to find p_m we notice that $q_2 = (1 - \alpha_2)p_1^{(0)}(x_1)p_2^{(0)}(x_2)$, hence $p_2 = p_1^{(0)}(x_1)p_2^{(0)}(x_2)$. Then, assuming that $p_{m-1} = p_1^{(0)}(x_1) \cdots p_{m-1}^{(0)}(x_{m-1})$ we see from (33), using (8), that $q_m = (1 - \alpha_m)p_1^{(0)}(x_1) \cdots p_m^{(0)}(x_m)$. (See also (31′) (31″), (31‴).) Thence

$$(37) \qquad \lim_{n\to\infty} p^{(n)}(x_1, x_2, \cdots, x_m) = p_1^{(0)}(x_1)p_2^{(0)}(x_2) \cdots p_m^{(0)}(x_m).$$

The last formula contains the limit theorem we wished to prove. It can be stated as follows:

THEOREM II. *If m characters are arbitrarily linked, with the one exception of "complete linkage", the distribution of transmitted genes $p^{(n)}(x_1, \cdots, x_m)$ "converges towards independence." The limit distribution is the product of the m marginal distributions of the first order $p_i^{(0)}(x_i)$, which are derived from $p^{(0)}(x_1, \cdots, x_m)$, the distribution of gametes in the initial generation.*

If, however, the initial distribution $p^{(0)}(x_1, \cdots, x_m)$ shows particular features, the stationary state may be reached already for a finite value of n. This happens with $n = 0$ and for every l.d. if $p^{(0)}(x_1, \cdots, x_m) = p_1^{(0)}(x_1) \cdots p_m^{(0)}(x_m)$. In other particular cases it may happen under particular assumptions for the l.d.

Let us express the general result also in terms of the distribution of genotypes. It follows from (37) and (27) that

$$\lim_{n\to\infty} w^{(n+1)}(x; y) = \lim_{n\to\infty} p^{(n)}(x)p^{(n)}(y)$$

$$= p_1^{(0)}(x_1) \cdots p_m^{(0)}(x_m)p_1^{(0)}(y_1) \cdots p_m^{(0)}(y_m) = \prod_{i=1}^{m} [p_i^{(0)}(x_i)p_i^{(0)}(y_i)].$$

Now consider a product like $p_i^{(0)}(x_i)p_i^{(0)}(y_i)$. By definition of $p_1^{(0)}(x_1)$ and applying (27) we find

$$p_1^{(0)}(x_1)p_1^{(0)}(y_1) = \sum_{x_2} \cdots \sum_{x_m} p^{(0)}(x_1, \cdots, x_m) \sum_{y_2} \cdots \sum_{y_m} p^{(0)}(y_1, \cdots, y_m)$$

$$= \sum_{x_2,\cdots,x_m} \sum_{y_2,\cdots,y_m} p^{(0)}(x)p^{(0)}(y)$$

Introducing then in a natural way the marginal distribution:

$$w_i^{(n)}(x_i ; y_i) = \sum_{x_1,\cdots,x_{i-1}x_{i+1},\cdots,x_m} \sum_{y_1,\cdots,y_{i-1}y_{i+1},\cdots,y_m} w^{(n)}(x_1, \cdots, x_m ; y_1, \cdots, y_m) \tag{38}$$

it is seen that

$$p_i^{(0)}(x_i)p_i^{(0)}(y_i) = w_i^{(1)}(x_i ; y_i). \tag{39}$$

Thence the result

$$\lim_{n\to\infty} w^{(n)}(x_1, \cdots, x_m ; y_1, \cdots, y_m) = w_1^{(1)}(x_1 ; y_1) \cdots w_m^{(1)}(x_m ; y_m) \tag{40}$$

which may be stated as follows:

THEOREM III. *In case of m arbitrarily linked Mendelian characters the distribution of the genotypes in the nth generation, $w^{(n)}(x_1, \cdots, x_m ; y_1, \cdots, y_m)$, "approaches independence" as $n \to \infty$. The limit distribution is the product of the m marginal distributions $w_i^{(1)}(x_i ; y_i)$ of the ith character ($i = 1, \cdots, m$) in the first filial generation.*

This theorem, which may be regarded as a corollary to THEOREM II, holds for any type of linkage, except "complete linkage" as defined in (17) where (36) is valid.

8. Solution of the recurrence equations (31). Formula (31) expresses $p^{(n)}(x_1, \cdots, x_m)$ in terms of $p^{(n-1)}(x_1, \cdots, x_m)$ (and all marginal distributions of $p^{(n-1)}$) and of the l.d. It seems desirable to try to express $p^{(n)}(x)$ in terms of $p^{(0)}(x)$. Now (31) is not a single equation but rather a complex system of difference equations with constant coefficients because for each marginal distribution of order $i < m$ the respective recurrence formula (31) of order i has to be used. (Or, if it is preferred to consider the marginal distributions as sums of p-values of order m, then all these p-values appear simultaneously and there is again a complicated system of difference equations.) In this situation it is not to be expected that the integration will yield simple explicit formulae, particularly as long as the l.d. is left arbitrary. However, the construction of the following formulae is clear. They reduce to simpler expressions in particular cases.

Let us use a method of indeterminate coefficients. To simplify the writing denote $p^{(0)}(x_1, \cdots, x_m)$ and its marginal distributions $p_i^{(0)}(x_i)$, $p_{ij}^{(0)}(x_i, x_j)$, etc. by $p_{12,\ldots,m}$, p_i, p_{ij}, etc. From genetical as well as mathematical considerations we gather the general form of $p_{12\cdots m}^{(n)}$ in terms of $p_{12\ldots m}$ and its marginal distributions; that this is indeed the general form will be verified by our very computations. Consider the set S consisting of the m numbers $1, 2, \cdots m$ and divide S in every possible way in two disjoint parts A_1 and A_2, none of them being empty, so that $A_1 + A_2 = S$, then divide S in every possible way into three disjoint parts so that $A_1 + A_2 + A_3 = S$, and finally S is divided into m disjoint parts each consisting of one single element. Denoting the unknown coefficients in a corresponding way by $\alpha_S^{(n)}$, $\alpha_{A_1,A_2}^{(n)}$, $\alpha_{A_1,A_2,A_3}^{(n)}$, etc. and writing $p_S^{(n)}$ and p_S for $p_{12\cdots m}^{(n)}$ and $p_{12\cdots m}^{(0)}$ the general form of p_S will be

$$p_S^{(n)} = \alpha_S^{(n)} p_S + \sum_{(A_1)} \alpha_{A_1,A_2}^{(n)} p_{A_1} p_{A_2} + \sum_{(A_1,A_2)} \alpha_{A_1,A_2,A_3}^{(n)} p_{A_1} p_{A_2} p_{A_3} + \cdots + \alpha_{1,2,3,\cdots,m}^{(n)} p_1 p_2 p_3 \cdots p_m. \tag{41}$$

This holds for every m. We get e.g. for $m = 4$

$$\begin{aligned} p_{1234}^{(n)} = {} & \alpha_{1234}^{(n)} p_{1234} + (\alpha_{1,234}^{(n)} p_1 p_{234} + \alpha_{2,134}^{(n)} p_2 p_{134} + \cdots) \\ & + (\alpha_{12,34}^{(n)} p_{12} p_{34} + \cdots) + (\alpha_{12,3,4}^{(n)} p_{12} p_3 p_4 + \cdots) + \alpha_{1,2,3,4}^{(n)} p_1 p_2 p_3 p_4. \end{aligned} \tag{41'}$$

For $m = 6$, e.g., there are eleven different types: One term $\alpha_{1\cdots 6}^{(n)} p_{1\ldots 6}$; then 6 terms of the form $\alpha_{1,2\cdots 6}^{(n)} p_1 p_{23\ldots 6}$; 15 terms like $\alpha_{12,3456}^{(n)} p_{12} p_{3456}$; 10 terms like $\alpha_{123,456}^{(n)} p_{123} p_{456}$; 15 terms like $\alpha_{1,2,3456}^{(n)} p_1 p_2 p_{3456}$; 60 terms like $\alpha_{1,23,456}^{(n)} p_1 p_{23} p_{456}$; 15 terms like $\alpha_{12,34,56}^{(n)} p_{12} p_{34} p_{56}$; 20 terms as $\alpha_{1,2,3,456}^{(n)} p_1 p_2 p_3 p_{456}$; 15 terms as $\alpha_{12,34,5,6}^{(n)} p_{12} p_{34} p_5 p_6$; 15 terms as $\alpha_{1,2,3,4,56}^{(n)} p_1 p_2 p_3 p_4 p_{56}$; and one final term $\alpha_{1,2,3,4,5,6}^{(n)} p_1 p_2 p_3 p_4 p_5 p_6$.

In (41) the $\alpha_{\cdots}^{(n)}$ are unknown constants depending on n and on the l.d. In order to find them consider (31) and write for the values of the l.d. v_A^m instead of $2l(A)$ (no confusion is possible because no marginal distribution of the l.d. occurs in (31)). With this notation (31″) e.g. reads:

$$p_{1234}^{(n+1)} = v_0^4 p_{1234}^{(n)} + (v_1^4 p_1 p_{234}^{(n)} + \cdots) + (v_{12}^4 p_{12}^{(n)} p_{34}^{(n)} + \cdots). \tag{31''}$$

If there is no ambiguity the upper m in v_A^m may even be omitted. Now assume the equations (41) to be written for $\mu = 2, \mu = 3, \cdots, \mu = m$. Introduce into the left side of (31) the expression (41) for $p_S^{(n+1)}$ and in the same way replace on the right side of (31) all $p_i^{(n)}, p_{ij}^{(n)}, \cdots, p_{12\cdots m}^{(n)}$ by their respective expressions (41). In this way an equality is obtained from which recurrence formulae for the unknown coefficients may be deduced by collecting all groups of terms which contain the same products of p's.

If this is carried out, e.g. for $m = 4$, the recurrence formulae are

$$\begin{aligned} \alpha_{1234}^{(n+1)} &= v_0\alpha_{1234}^{(n)} \\ \alpha_{123,4}^{(n+1)} &= v_0\alpha_{123,4}^{(n)} + v_4\alpha_{123}^{(n)} \\ \alpha_{12,34}^{(n+1)} &= v_0\alpha_{12,34}^{(n)} + v_{12}\alpha_{12}^{(n)}\,\alpha_{34}^{(n)} \\ \alpha_{12,3,4}^{(n+1)} &= v_0\alpha_{12,3,4}^{(n)} + v_{12}\alpha_{12}^{(n)}\alpha_{3,4}^{(n)} + v_3\alpha_{12,4}^{(n)} + v_4\alpha_{12,3}^{(n)} \\ \alpha_{1,2,3,4}^{(n+1)} &= v_0\alpha_{1,2,3,4}^{(n)} + v_1\alpha_{2,3,4}^{(n)} + \cdots + v_{12}\alpha_{1,2}^{(n)}\alpha_{3,4}^{(n)} + \cdots \end{aligned} \quad \text{etc.} \tag{42}$$

In general, i.e. for any m, these recurrence formulae are of a clear structure the first one being particularly simple, namely

$$\alpha_S^{(n+1)} = v_0\alpha_S^{(n)}. \tag{43}$$

It can be solved immediately and gives

$$\alpha_S^{(n)} = v_0^n\,. \tag{43'}$$

The other recurrence formulae are all of the form

$$x_{n+1} = v_0x_n + f(n) \text{ with } x_0 = 0, \tag{44}$$

where $f(n)$ is a given function of n whose general form is still to be investigated. The solution of (44) is

$$x_n \equiv \sum_{\nu=0}^{n-1} f(\nu)v_0^{n-1-\nu}. \tag{44'}$$

With the notations used in (41) the equation (44) may be written:

$$\alpha_{A_1,A_2,\cdots,A_\mu}^{(n+1)} = v_0\alpha_{A_1,A_2,\cdots,A_\mu}^{(n)} + A_{A_1,A_2,\cdots,A_\mu}^{(n)}\,. \tag{44''}$$

We have to determine $A_{A_1,A_2,\ldots,A_\mu}$. For reasons of symmetry and homogeneity let us introduce constants $\alpha_1^{(n)} = \alpha_2^{(n)} = \cdots = \alpha_m^{(n)} = 1$. With that notation e.g. the last term in the second line in (42) reads $v_4\alpha_{123}^{(n)}\alpha_4^{(n)}$ or the third term to the right in the fourth line of (42): $v_3\alpha_{12,4}^{(n)}\alpha_3^{(n)}$ etc.

The construction of $A_{A_1,A_2,\ldots,A_\mu}^n$ may then be described as follows: Each $A_{A_1,A_2,\ldots,A_\mu}^n$ is a sum of $2^{\mu-1} - 1$ terms, each term being a product of one v-value and two α's. The set consisting of the μ elements $A_1, A_2, \cdots, A_\mu$ is to be divided in all possible ways into two non-empty, disjoint, complementary parts which form the subscripts of the two α's in question; the subscript of v is equal to the subscript of either of these two α-values; it makes no difference which, because of the specific symmetry (8′) of the l.d.; it should be noted that in the subscripts of v no comma occurs. As an example let us write $A^{(n)}_{1234,567,8}$ for $m = 8$. We get: $A^{(n)}_{1234,567,8} = v_8\alpha_8^{(n)}\alpha^{(n)}_{1234,567} + v_{567}\alpha^{(n)}_{567}\alpha^{(n)}_{1234,8} + v_{1254}\alpha^{(n)}_{1234}\alpha^{(n)}_{567,8}$. Or if we wish $A^{(n)}_{12,34,5,6}$ for $m = 6$: $A^{(n)}_{12,34,5,6} = v_6\alpha_6^{(n)}\alpha^{(n)}_{12,34,5} + v_5\alpha_5^{(n)}\alpha^{(n)}_{12,34,6} + v_{12}\alpha^{(n)}_{12}\alpha^{(n)}_{34,5,6} + v_{34}\alpha^{(n)}_{34}\alpha^{(n)}_{12,5,6} + v_{56}\alpha^{(n)}_{5,6}\alpha^{(n)}_{12,34} + v_{125}\alpha^{(n)}_{12,5}\alpha^{(n)}_{34,6} + v_{126}\alpha^{(n)}_{12,6}\alpha^{(n)}_{34,5}$.

Hence, in principle our "integration" problem, where n is the variable, is completely solved: First $p_S^{(n)}$ is given by (41). Then, in order to find any $\alpha^{(n)}_{A_1,A_2,\ldots A_\mu}$, we first determine the corresponding $A_{A_1,A_2,\ldots A_\mu}$ by the rule just explained and illustrated, and then it follows from (44′) that

$$(44''') \qquad \alpha^{(n)}_{A_1,A_2,\cdots,A_\mu} = \sum_{\nu=0}^{n-1} v_0^{n-1-\nu} A^{(\nu)}_{A_1,A_2,\cdots,A_\mu} .$$

This whole procedure, although in principle very simple, may of course be lengthy if m is not small and if no specific assumption for the l.d. is considered; for in the expression of $A_{A_1,A_2,\ldots,A_\mu}$ many different α-values appear,—each however with less than m subscripts—which play the role of abbreviations for complicated expressions; in other words the explicit solution for $m = 6$, for instance requires the solutions for $m < 6$, all these solutions being however completely given by our formulae, down to $m = 2$, where $\alpha^{(n)}_{12}$ and $\alpha^{(n)}_{1,2}$ are given by (23).

Under simple assumptions for the l.d. the explicit expressions for the α become simple. Two extreme cases are complete linkage and free assortment. In the first case $p^{(n)}_{12\cdots m} = p^{(0)}_{12\cdots m}$ and nothing remains to be done. The case of free assortment where all $v = (\frac{1}{2})^{m-1}$ can be dealt with directly by induction, or we may evaluate the general formulae given above which in this case become quite simple. We have[8]

$$(45) \qquad 2^{mn}\alpha^{(n)}_{A_1,A_2,\ldots,A_\mu} = 2^n(2^n - 1)\cdots(2^n - \mu + 1).$$

That shows that the values of the coefficients $\alpha^{(n)}_{\cdots}$ depend only on the number of elements A_i which appear as subscripts. Thus we find e.g. for $m = 6$, if we write in each line of (45′) one typical value:

$$(45') \qquad \begin{aligned}
\alpha^{(n)}_{123456} &= 2^n/2^{6n} = 1/2^{5n} \\
\alpha^{(n)}_{1,23456} &= \alpha^{(n)}_{12,3456} = \alpha^{(n)}_{123,456} = (2^n - 1)/2^{5n} \\
\alpha^{(n)}_{1,2,3456} &= \alpha^{(n)}_{1,23,456} = \alpha^{(n)}_{12,34,56} = (2^n - 1)(2^n - 2)/2^{5n} \\
\alpha^{(n)}_{1,2,3,456} &= \alpha^{(n)}_{12,34,5,6} = (2^n - 1)(2^n - 2)(2^n - 3)/2^{5n} \\
\alpha^{(n)}_{1,2,3,4,56} &= (2^n - 1)(2^n - 2)\cdots(2^n - 4)/2^{5n} \\
\alpha^{(n)}_{1,2,3,4,5,6} &= (2^n - 1)(2^n - 2)\cdots(2^n - 5)/2^{5n} .
\end{aligned}$$

Thus in the simple case of independent assortment the explicit solution is very simple too. It confirms the fact that $\lim_{n\to\infty} \alpha^{(n)}_{1,2,3,4,5,6} = 1$ while all other α's approach

[8] The values on the right side of (45) are indicated in [1]; but the solution for free assortment reported in this article does not seem to coincide with ours.

zero. To prove this, however, without recurring to computations, was the purpose of the preceding section.

9. Crossover distribution and crossover probabilities. The limit theorem of §7 as well as the computations of the preceding section, in short, all investigations and concepts considered so far, are valid for any l.d. We shall now define and use a *crossover distribution*, (c.d.), which is completely equivalent to the l.d. but preferable for the study of certain particular cases. Apparently biologists have not considered the general concept of the c.d. but only the c.p. c_{ij}. This concept is basic and tangible but not sufficient for a complete description of the linkage mechanism when $m \geqq 4$, as was seen in the preceding sections.

It is obvious that, from our point of view, a mathematical theory of linkage must be based on the properties of and a set of assumptions on the l.d., or the c.d. The *linear theory* will be considered from this standpoint. This theory is, of course, still compatible with a variety of particular assumptions. In the last section some simple particular cases will be presented and studied with a special view to *interference.*

The probability that an individual transmits the set of "paternal genes" belonging to A and the set of "maternal genes" belonging to $A' = S - A$ is denoted by $l(A)$, where $l(A) = l(A')$; e.g. with $m = 8$: $l(1,0,1,1,1,0,0,1) = l(0,1,0,0,0,1,1,0)$. Considering here the succession of arguments we see that in either set of eight arguments: The first and the second are from different sets, the second and the third are again from different sets, the third and the fourth are from the same set, $\cdots$ the seventh and eighth are from different sets. Writing 0 for "same" and 1 for "different" and using these numbers to correspond to the $(m - 1)$ consecutive intervals between the m genes, we introduce:

$$l(10111001) + l(01000110) = \pi(1100101).$$

Here $\pi(\eta_1, \eta_2, \cdots, \eta_{m-1})$ where $\eta_i = 0$ or 1, is an $(m - 1)$-variate alternative. The relation between the l.d. and this new distribution may be written in the form

$$(46) \qquad 2l(\epsilon_1, \epsilon_2, \cdots, \epsilon_m) = \pi(|\epsilon_1 - \epsilon_2|, |\epsilon_2 - \epsilon_3|, \cdots |\epsilon_{m-1} - \epsilon_m|), \quad \epsilon_i = 0 \text{ or } 1.$$

In this definition no fixed "order" of the genes is implied so far. The numbers $1, 2 \cdots m$ are used like names.

But it seems to be admitted today by leading biologists that a certain natural order of the genes exists. If this is so the numbers $1, 2, \cdots m$ should be used in agreement with this order. Let us note, however, that the situation is in reality slightly different: Only the genes *within each linkage group* (§4) are assumed to be ordered, whereas no order exists among the groups. Let us for the moment disregard this circumstance and assume that all genes under consideration belong to the same linkage group.

Within such a linkage group a one-dimensional or linear order prevails, to be understood in the geometric sense of "location". Some more precise definitions concerning this linear order will be considered later. For the moment we simply imagine that each of the two sets of genes belonging to an individual is arranged like m consecutive discrete points on a line segment.[9] The crossover distribution $\pi(\eta_1, \eta_2, \cdots \eta_{m-1})$, introduced in (46) becomes more meaningful under this assumption where, now, the numbering corresponds to this linear order. Then the argument 0 in this distribution can be interpreted as "coherence" and the argument 1 as "interchange" or "crossing over" and the "intervals" as intervals in the geometric sense. Whether this "crossing over", which means transition from the maternal to the paternal set or vice versa, is to be conceived as a "break" (Janssen's chiasmatypie) does not matter for the above definitions. If however, the idea is that between two neighboring genes not more than one break is possible then the "event," which we call crossover, would be at the same time a break; if, biologically, more than one break between i and $(i + 1)$ is not excluded, then the event "crossover within $(i, i + 1)$" means "odd number of breaks within this interval."

Now, let us consider the relation between the c.d. and the c.p. It has been repeatedly remarked that the c.p. are not equivalent to the l.d., hence they are not equivalent to the c.d. either. There are $\frac{1}{2} \cdot m(m - 1)$ c.p. but $2^{m-1} - 1$ l-values, or π-values. If $m \geqq 4$ the second number is greater than the first. Besides, the l-values are absolutely arbitrary probabilities. For the c.p. in section 4 some restrictions were derived. Let us derive another *set of restrictions* by considering four numbers i, j, k, l which we may denote by 1, 2, 3, 4. (The following computation has nothing to do with linear order. It applies *if* $m = 4$ to the l.d. $l(\epsilon_1\epsilon_2\epsilon_3\epsilon_4)$ and if $m > 4$ to the respective four-dimensional marginal distributions of the l.d.) Write $v(\epsilon_1\epsilon_2\epsilon_3\epsilon_4) = 2l(\epsilon_1\epsilon_2\epsilon_3\epsilon_4)$ and let us add up the six c.p. corresponding to these four numbers. From $c_{ij} = 2l_{ij}(1,0) = v_{ij}(1,0)$ we get

$$\begin{aligned} c_{12} + c_{13} + \cdots + c_{34} &= 3v(1000) + 3v(0100) + 3v(0010) + 3v(0001) \\ &\quad + 4v(1001) + 4v(1010) + 4v(1100) \\ &= 4 - 4v(0000) - v(1000) - v(0100) - v(0010) - v(0001) \leqq 4. \end{aligned} \tag{47}$$

Hence as by (14) $c_{12} + c_{23} + c_{13} \leqq 2$, it follows that

$$c_{il} + c_{jl} + c_{kl} \leqq 2 \tag{14'}$$

is another necessary condition for the c.p. The limit "2" can be reached, as we see for $v(0000) = v(1000) = v(0100) = v(0010) = v(0001) = 0$; then

$$c_{12} = c_{34} = v(1001) + v(1010)$$

$$c_{23} = c_{14} = v(1100) + v(1010)$$

$$c_{13} = c_{24} = v(1001) + v(1100)$$

[9] "The genes are represented as lying in a line like beads on a string. The numerical data from crossing over show in fact that this arrangement is the only one that is consistent with the results obtained" [11]. This is but one of many statements in favor of the linear theory.

and

$$c_{12} + c_{23} + c_{13} = c_{14} + c_{24} + c_{34} = 2.$$

To summarize the facts about the c.p.: *In case of m characters there are* $\frac{1}{2}m(m - 1)$ c.p. $c_{ij} = 2l_{ij}(10) = 2l_{ij}\,(0,1)$. *These values must satisfy the following necessary conditions* (besides $0 \leqq c_{ij} \leqq 1$):

(13) $$c_{ij} + c_{jk} \geqq c_{ik} \qquad \text{for any three subscripts}$$

(14) $$c_{ij} + c_{jk} + c_{ik} \leqq 2 \qquad \text{“ \quad “ \quad “ \quad “}$$

(14′) $$c_{il} + c_{jl} + c_{kl} \leqq 2 \qquad \text{“ \quad “ \ four \quad “}$$

If in an analogous way five or more subscripts are considered no new condition turns up. It has, however, not been proved that the above given necessary conditions are sufficient for a consistent system of c.p. If we wish to be sure of consistency the starting point must be a l.d. or a c.d. from which the c_{ij} are deduced.

[This question of consistency belongs in the same class as the following problem: "Under what conditions does a set of $\binom{m}{2}$ distributions $V_{ij}(x_i\,, x_j)$ form the marginal distributions of second order of an m-dimensional distribution $V(x_1\,, \cdots, x_m)$?" Here $V(x_1, \cdots, x_m)$ is the probability that the first result is $\leqq x_1$, the second $\leqq x_2$, the last $\leqq x_m$. An analogous question arises for the set of $\binom{m}{3}$ distributions $V_{ijk}(x_i\,, x_j\,, x_k)$, etc.[10]]

In the following it will be necessary to know *the expressions of the c.p. in terms of the c.d.* Put $m - 1 = n$ and denote by p_i, p_{ij}, etc. in the usual way the following probabilities derived from the c.d.: p_i is the probability of "success" in the i-th trial, p_{ij} the probability of success in both the i-th and j-th trial, etc. It has to be kept in mind that for the c.d. and all magnitudes derived from it the "i-th trial" is associated with the i-th interval, i.e. with the interval $(i, i + 1)$ "and success in the i-th trial" means cross over in this interval. [Whereas in the l.d. and in magnitudes derived from it, like $c_{ij} \equiv l_{ij}\,(1, 0)$ the subscript i denotes the i-th gene. (See (46)).] Now denote by S_1 the sum of all probabilities p_i, by S_2 the sum of all p_{ij}, $\cdots$. Besides, let $P_{1,\ldots,i}(x)$ be the probability of exactly x successes in the first i trials ($i = 1,2, \cdots, n$), and analogously, $P_{2\ldots j}(x)$ the probability of x successes in the $j - 1$ trials $2, 3, \cdots, j$, etc. Then the desired formulae follow easily: First we have obviously

$$c_{i,i+1} = p_i \qquad (i = 1,2, \cdots, n).$$

Because $c_{i,i+1}$ is the probability of one interchange between the genes i and $i + 1$ i.e. of an interchange in the i-th interval, of "success in the i-th trial".

[10] For one-dimensional distributions $V_i(x)$ the question is trivial because any set of m distributions $V_i(x)$ can be considered as the marginal distributions of first order of $V(x_1\,, \cdots, x_m) = V_1(x_1) \cdots V_m(x_m)$.

Then $c_{i,i+2}$ is the probability of one interchange between i and $i + 2$, i.e. of either an interchange in the first of the two intervals numbered i and $i + 1$, and no interchange in the second; or of an interchange in the second but none in the first. Hence $c_{i,i+2} = P_{i,i+1}(1)$, $(i = 1, \cdots, n - 1)$, because $P_{i,i+1}(1)$ is just the probability of exactly one "success" in the two trials numbered i and $i + 1$. In the same way we get $c_{i,i+3} = P_{i,\ldots,i+2}(1) + P_{i,\ldots,i+2}(3)$, $(i = 1, \cdots, n - 2)$, because an interchange between i and $i + 3$ means either exactly one or exactly three interchanges in the three intermediate intervals. Hence we get altogether, with $n = m - 1$:

$$\begin{aligned} c_{i,i+1} &= p_i \\ c_{i,i+2} &= P_{i,i+2}(1) \qquad (48) \\ c_{1m} &= P_{12\ldots n}(1) + P_{12\ldots n}(3) + \cdots P_{12\cdots n}(\bar{n}), \text{ where } \bar{n} = n \text{ if } n \text{ odd,} \\ &\qquad\qquad = n - 1 \text{ if } n \text{ even.} \end{aligned}$$

Let us also express the c_{ij} in terms of the S_i. It is well known (see e.g. [3]) that

$$P_{1,\cdots,n}(x) = \sum_{\nu=x}^{n} (-1)^{\nu+x} S_\nu \qquad (x = 0, 1, \cdots n). \qquad (49)$$

Applying these to (48) we easily find the convenient expressions:

$$\begin{aligned} c_{12} &= p_1, \text{ etc.} \\ c_{13} &= (p_1 + p_2) - 2p_{12} \equiv (S_1 - 2S_2)_{12}, \text{ etc.} \\ c_{14} &= (p_1 + p_2 + p_3) - 2(p_{12} + p_{13} + p_{23}) + 4p_{123} \\ &\equiv (S_1 - 2S_2 + 4S_3)_{123}, \text{ etc.} \qquad (50) \\ c_{15} &= (S_1 - 2S_2 + 4S_5 - 8S_4)_{1\ldots4}, \text{ etc.} \\ &\cdots\cdot \\ c_{1,m} &= S_1 - 2S_2 + 4S_3 + \cdots + (-2)^m S_{m-1}. \end{aligned}$$

10. The linear theory. Consider a linkage group of size m and assume for the moment that $c_{ij} \neq c_{ik}$ for all i, j, and k. It seems that the main mathematical content of the linear theory can be summarized as follows: *It is possible to establish in a unique way an order or a succession of the genes, such that for the* $\binom{m}{2} = \frac{m(m-1)}{2}$ *c.p. the* $(m - 1)(m - 2)$ *inequalities*

$$\begin{aligned} c_{ij} &< c_{i,j+1} \qquad (i = 1, 2, \cdots, m - 2) \\ c_{ij} &< c_{i-1,j} \qquad (i = 2, 3, \cdots, m - 1) \end{aligned} \qquad (i < j) \qquad (51)$$

hold. In this succession j will be between i and k if c_{ik} is greater than the two other c.p. c_{ij} and c_{jk}. The two arrangements $1, 2, \cdots m$ and $m, m - 1, \cdots,$

$\cdots$ 1 are considered as corresponding to the same order. *Furthermore, this order is a straight-line-succession for which an additive distance relation holds* (cf. also [4a] and [13]). Instead of the restriction $c_{ij} \neq c_{ik}$ it is sufficient to assume *the weaker restriction, that in any triple c_{ij}, c_{ik}, c_{kj} one is greater than the two others.* Without such a restriction uniqueness of the order no longer holds. E.g. in case of independent assortment where all c_{ij} equal $\frac{1}{2}$ any of the m: possible numberings of the genes is equally admissible from the point of view of the linear theory. In the case of complete linkage where all c.p. are zero it will be logical to consider all m genes as located in the same point. Obviously there are all kinds of intermediate cases. We shall come back to this point at the end of this section.

Now consider again the case of "*different*" c.p. (in the above defined sense). Let us prove that there can be *not more than one succession* for which (51) holds. In fact it follows from (51) that also:

$$\begin{array}{llll} (51') & c_{ij} < c_{ik} & (i = 1, 2, \cdots, m-2) & \text{for all} \quad k > j \\ \text{and} & c_{ij} < c_{kj} & (i = 2, 3, \cdots, m-1) & \text{for all} \quad k < i. \end{array} \qquad i < j$$

These are all together $M = 2\cdot 1 + 3\cdot 2 + \cdots + (m-1)(m-2) = 2\cdot\binom{m}{3}$ inequalities. On the other hand there are all together $\binom{m}{3} = M/2$ "between"-relations for m numbers, each of them being defined by two inequalities as $c_{ij} < c_{ik}$ and $c_{jk} < c_{ik}$ (if j is between i and k); hence on the whole M such inequalities. But these are the same as (51′), as we see by changing i, j, k into j, k, i in the second equation (51′). Thus it is not possible to find two different successions which both satisfy (51).

As to the *metric* of the problem, Morgan proposed originally that the value of the c.p. c_{ij} should be used as the distance between i and j. It has, however been objected repeatedly that this distance would not be additive; this is obvious since the triangular relation (13) holds for three subscripts (see also (50).[11] The equality $c_{ij} + c_{jk} = c_{ik}$ holds only in the exceptional cases where multiple crossingover is excluded. It seems, however that an adequate definition of distance is available if we try to formulate in terms of probability theory what the biologist had in mind. Let $j \geqq i$. *The distance $d_{i,j+1}$ between i and $j + 1$ may be defined as the mathematical expectation of the number of crossingovers in $(i, j + 1)$,* i.e. in the $j + 1 - i$ intervals between i and $j + 1$. Hence if

[11] For a geometric equivalent of m points with $m(m-1)/2$ *arbitrary* distances we would have to turn to an $(m-1)$-dimensional space. In fact it is well known that there are between k points in the plane only $S_2 = 2k - 3$ arbitrary distances, in space only $S_3 = 3k - 6$, in r-space $S_r = rk - r\,(r+1)/2$. Hence for $r = m - 1$ and $k = m$: $S_{m-1} = m(m-1)/2$.

$P_{i,\ldots,j}(x)$ denotes, as before, the probability of exactly x crossovers in these $(j + 1 - i)$ intermediate intervals the formula holds

$$d_{i,j+1} = \sum_{x=0}^{j+1-i} xP_{i\cdots j}(x). \tag{52}$$

Of course, an appropriate unit may be used such that in practical use the distance becomes *proportional* to the d_{ij} introduced above.

The mean value to the right in (52) is well known for any distribution $\pi(x_1, \cdots, x_n)$ whether an "independent" or a general distribution; (i.e. in our case: with or without "interference"). Denoting in the usual way by $\pi_i(x_i)$ the marginal distributions of first order of $\pi(x_1, \cdots, x_n)$ and putting $\pi_i(1) = p_i =$ the probability of success in the i-th trial, we get:

$$d_{i,j+1} = p_i + p_{i+1} + \cdots + p_j \tag{53}$$

and in the same way with $k > j$

$$\begin{aligned} d_{j+1,k+1} &= p_{j+1} + p_{j+2} + \cdots + p_k \\ d_{i,k+1} &= p_i + p_{i+1} + \cdots + p_k \end{aligned}$$

hence $d_{i,j+1} + d_{j+1,k+1} = d_{i,k+1}$, or in general:

$$d_{ij} + d_{jk} = d_{ik} \qquad (i < j < k). \tag{54}$$

It may be mentioned that the additive property of the mathematical expectation which was used here is very well known (particularly for independent events) but not always correctly proved. The proof is contained in the transformation expressed in the following equalities:

$$\begin{aligned} d_{i,j+1} &\equiv \sum_{x=0}^{j+1-i} xP_{i\cdots j}(x) \\ &= \sum_{x_i} \sum_{x_{i+1}} \cdots \sum_{x_j} (x_i + x_{i+1} + \cdots + x_j)\pi_{i,i+1,\cdots j}(x_i, \cdots, x_j) \\ &= \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} (x_i + x_{i+1} + \cdots + x_j)\pi(x_1, x_2, \cdots, x_n). \end{aligned} \tag{55}$$

(For general distributions Stieltjes integrals replace the sums.) In (55) $\pi(x_1, \cdots, x_n)$ is the given n-variate distribution, $P_{i,\ldots,j}$ the probability of exactly x successes in the successive trials numbered $i, \cdots, j$ and $\pi_{i,i+1,\ldots,j}(x_i, \cdots x_j)$ is the respective marginal distribution of $\pi(x_1, \cdots, x_n)$. The first equality in (55) is not obvious, while the second is rather trivial. From the second or third form of $d_{i,j+1}$ in (55), follows (53). The last expression in (55) shows that the expectation of any such sum as $(x_i + x_{i+1} + \cdots x_j)$ can be computed with respect to one and the same distribution $\pi(x_1, \cdots, x_n)$. Therefore *the distance $d_{i,j+1}$ may also be defined as the expectation of $(x_i + x_{i+1} + \cdots + x_j)$ with respect to the c.d.*

Because of the first equation (48) we get from (53)

$$d_{ij} = c_{i,i+1} + c_{i+1,i+2} + \cdots + c_{j-1,j}. \tag{53'}$$

Hence the distance d_{ij} is equal to the sum of the $j - i$ intermediate c.p. No difficulty arises for us from the obvious fact that always

$$c_{ij} \leqq d_{ij}; \text{ and in general } c_{ij} < d_{ij}, \tag{53''}$$

because the distance d_{ij} is defined by (52), or (55) and not as c_{ij}.

On the right side in (53′) stands the sum of certain c.p. We have repeatedly remarked that there may be hitherto unknown restrictions for a consistent system of c.p. Hence it is important to notice that *there are no restrictions for the particular $(m - 1)$ c.p. $c_{12}, c_{23}, \cdots, c_{m-1,m}$. They can be quite arbitrarily chosen* because of $c_{i,i+1} = p_i$. Hence *any geometric representation of m genes arranged on a straight line in arbitrary distances $d_{i,i+1}$ $(i = 1, 2, \cdots m - 1)$ is surely consistent.* E.g. m consecutive genes may be arranged with equal distances $d_{12} = d_{23} = \cdots = d_{m-1,m}$. Or some distances may be zero; then the respective genes are localized in the same point, etc.

Finally, let us briefly consider the case of *several linkage groups.* According to §4 the l.d. then resolves into a product of several distributions; e.g. with $m = 12$:

$$\begin{aligned} l(\epsilon_1\epsilon_2, \cdots, \epsilon_{12}) &= f_1(\epsilon_1\epsilon_2\epsilon_3\epsilon_4)f_2(\epsilon_5\epsilon_6\epsilon_7)f_3(\epsilon_8\epsilon_9\epsilon_{10})f_4(\epsilon_{11}\epsilon_{12}) \\ &= (\tfrac{1}{2})^4\pi_1(|\epsilon_1 - \epsilon_2|, |\epsilon_2 - \epsilon_3|, |\epsilon_3 - \epsilon_4|) \\ &\quad \pi_2(|\epsilon_5 - \epsilon_6|, |\epsilon_6 - \epsilon_7|) \cdots \pi_4(|\epsilon_{11} - \epsilon_{12}|). \end{aligned} \tag{56}$$

Then, as postulated by Morgan, *the linear order holds within each of the k groups, whereas all c.p. among the groups are equal to $\frac{1}{2}$.*

Let us conclude this section by transforming the basic conditions (51) of the linear theory by means of (48). This will be needed in the following section. Consider e.g. the condition $c_{13} < c_{14}$, i.e.

$$P_{12}(1) < P_{123}(1) + P_{123}(3) \tag{57'}$$

or $c_{24} < c_{14}$ yields $P_{23}(1) < P_{123}(1) + P_{123}(3)$. Or in the same way:

$$\begin{aligned} P_{123}(1) + P_{123}(3) &< P_{1234}(1) + P_{1234}(3) \\ &< P_{1,\ldots,5}(1) + P_{1,\ldots,5}(3) + P_{1,\ldots,5}(5), \text{ etc.} \end{aligned} \tag{57''}$$

Thus we may express the content of (51) as follows: *The probability that the "event" happens an odd number of times in a set T_i of i consecutive trials is less than the probability that the event happens an odd number of times in the set T_{i+1} or in the set T'_{i+1} each consisting of $i + 1$ consecutive trials where T_{i+1} and T'_{i+1} denotes respectively the sum of T_i and either the immediately following or the immediately preceding trial.* In this form we see again that the linear theory is an assumption, suggested by observations, and by no means logically necessary.

11. Some models of c.d.'s based on the linear theory. The simplest and very important example which has been suggested repeatedly is that of independent crossovers:

(*i*) *Independence.* The crossovers do not influence each other, i.e.

$$p_{ij} = p_i p_j, \qquad p_{ijk} = p_i p_j p_k, \cdots. \tag{58}$$

That this distribution is consistent is well known; hence only the specific inequalities (48) or (57) have to be considered. Here the expressions $P_{12\ldots i}(x)$, used in (57) become very simple, e.g. with $p_i + q_i = 1$:

$$P_{1\ldots 4}(1) = p_1q_2q_3q_4 + q_1p_2q_3q_4 + q_1q_2p_3q_4 + q_1q_2q_3p_4.$$

Then a simple computation shows:

$$\begin{aligned} c_{i,j+1} - c_{ij} &= (q_i - p_i) \cdots (q_{j-1} - p_{j-1})p_j \\ c_{i-1,j} - c_{ij} &= p_{i-1}(q_i - p_i) \cdots (q_{j-1} - p_{j-1}). \end{aligned} \tag{59}$$

These differences will be positive if all $q_i - p_i > 0$ or all $p_i < \frac{1}{2}$. Hence: *A consistent c.d. which fulfils the conditions* (51) *of the linear theory is the distribution of "independent crossovers" with basic probabilities* $p_i = c_{i,i+1}$ $(i = 1, 2, \cdots m - 1)$, *with the one restriction*

$$c_{i,i+1} = p_i \leqq \tfrac{1}{2}. \tag{60}$$

The distribution is completely determined by (58). If all $p_i = p = \frac{1}{2}$, we have the particular case of free assortment.

Although this independence is more general than Mendel's original assumption, Morgan, Haldane and others reported observations, not in accordance with this hypothesis. One crossingover seems to prevent others in a certain "neighborhood". This phenomenon was named *interference.* It suggests that we have to consider the c.d. as a distribution of dependent rather than of independent events. This will be done in the following pages. First consider the limit-case of:

(*ii*) *Complete interference or disjoint events.* In this case we have

$$p_{ij} = p_{ijk} = \cdots = p_{12,\ldots,m-1} = 0. \tag{61}$$

Thence it follows that we have simply

$$\begin{aligned} c_{i,i+1} &= p_i \\ c_{i,i+2} &= p_i + p_{i+1} \\ c_{i,i+3} &= p_i + p_{i+1} + p_{i+2}, \text{ etc.} \end{aligned} \tag{62}$$

In this particular case the c.p. are additive $c_{ij} = d_{ij}$. It is obvious from (62) that *in this case the conditions* (51) *of the linear theory are fulfilled.* On the other hand it follows from (49), (for $x = 0$) and (61) that *the system is consistent if and only if*

$$S_1 \equiv p_1 + p_2 + \cdots + p_n \leqq 1 \qquad (n = m - 1). \tag{63}$$

This is in accordance with the fact that nearly or exactly additive c.p. have been observed always in connection with very small p_i-values.

The most striking observation leading to the concept of interference was that $p_{ij} \leqq p_i p_j$, i.e. that double crossovers appeared less frequently than one would have assumed for independent crossovers, but that nevertheless they did appear sometimes. A particularly simple model of dependence or interference which starts with this fact, preserving however, the main structure of independence, is the following:

(*iii*) *One-parametric model of partial interference.* Assume as before $(m - 1)$ basic probabilities p_i and put

$$p_{ij} = \epsilon p_i p_j, \qquad p_{ijk} = \epsilon p_i p_j p_k, \cdots, \text{etc.} \qquad (0 \leqq \epsilon \leqq 1). \tag{64}$$

There is independence if $\epsilon = 1$, complete interference for $\epsilon = 0$ and partial interference for intermediate values of ϵ. Let us first investigate conditions for the consistency of this distribution. Necessary and sufficient conditions for a consistent distribution of arbitrarily linked events are well known (see e.g. [3] (b) p. 239). Write $m - 1 = n$. A system of $p_i, p_{ij}, \cdots, p_{1\ldots n}$ is consistent if it is possible to compute from these $(2^n - 1)$ values, 2^n *non-negative values* $\pi(\eta_1, \eta_2, \cdots \eta_n)$ $(\eta_i = 0$ or $1)$ *which have the sum one* and are given by the formulae:

$$\begin{aligned}
&\pi(11\cdots1) = p_{12\cdots n}\\
&\pi(11\cdots10) = p_{12,\cdots,(n-1)} - p_{12,\cdots,n}\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\pi(110\cdots0) = p_{12} - \sum_{(n_1)} p_{12n_1} + \sum_{n_1}\sum_{n_2} p_{12n,n_2} - \cdots \pm p_{12\cdots n}\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&\pi(00\cdots0) = 1 - \sum_{n_1} p_{n_1} + \sum_{n_1}\sum_{n_2} p_{n_1 n_2} - \cdots \pm p_{12\cdots n}.
\end{aligned} \tag{65}$$

Because of the symmetry of (64) it will be sufficient to check (65) by means of the relations (49) which can be obtained from (65) by collecting groups of equations such that the corresponding $\pi(\eta_1, \cdots, \eta_n)$ show all the same number of 1's as arguments. Write $P_{1\ldots n}(x) = P_x$ for the probability of x successes in the n trials where independent events with basic probabilities p_i are considered, and P'_x for the analogous probability corresponding to the distribution (64) and introduce in the same way S_i and S'_i where as before, $S_2 = \Sigma p_{ij}$, $S_3 = \Sigma p_{ijk}$, etc. We then find:

$$\begin{aligned}
P'_0 &= 1 - S_1 + \epsilon S_2 - \epsilon S_3 + \cdots\\
&= P_0 + (1 - \epsilon)(- S_2 + S_3 - S_4 + \cdots) = P_0\epsilon + (1 - S_1)(1 - \epsilon).
\end{aligned} \tag{66}$$

It follows that the expression $P_0\epsilon + (1 - S_1)(1 - \epsilon)$ must be $\geqq 0$. For $\epsilon = 0$ this condition reduces to (63), whereas for $\epsilon = 1$, there is no restriction at all. On the other hand this is the only restriction of this kind, because we find

$$\begin{aligned}
P'_1 &= S'_1 - 2S'_2 + 3S'_3 - 4S'_4 + \cdots = S_1 - 2\epsilon S_2 + 3\epsilon S_3 - 4\epsilon S_4 + \cdots\\
&= P_1 + (1 - \epsilon)(2S_2 - 3S_3 + 4S_4 \cdots) = P_1\epsilon + S_1(1 - \epsilon).
\end{aligned}$$

This last expression is always $\geqq 0$ because of $P_1 \geqq 0$, $S_1 \geqq 0$, $\epsilon \leqq 1$. Furthermore we find for $i \geqq 2$ that $P'_i = P_i\epsilon$, hence always non negative. Therefore our system is consistent under the one condition (66).

The additional restrictions corresponding to the linear theory have still to be considered. A simple computation yields the result

$$\begin{aligned} c_{i,j+1} - c_{ij} &= (1 - 2\epsilon p_i)(1 - 2\epsilon p_{i+1}) \cdots (1 - 2\epsilon p_{j-1})p_j \\ c_{i,j+1} - c_{i+1,j+1} &= p_i(1 - 2\epsilon p_{i+1}) \cdots (1 - 2\epsilon p_j). \end{aligned} \tag{67}$$

These differences are $\geqq 0$ if $p_i \leqq \frac{1}{2\epsilon}$ which is, for $\epsilon < 1$, less strong than (60). Hence we sum up: *A consistent model of partial interference with one parameter* ϵ *to fit the observations can be obtained on the basis of* $n = m - 1$ *probabilities* $p_1, p_2, \cdots p_n$ *by means of* (64), *if the condition*

$$P_0\epsilon + (1 - S_1)(1 - \epsilon) \geqq 0 \quad \text{or} \quad S_1 \leqq 1 + \frac{\epsilon}{1 - \epsilon} P_0 \tag{68}$$

holds and the additional restriction required by the "linear theory"

$$p_i \leqq \frac{1}{2\epsilon} \tag{69}$$

is satisfied. For $\epsilon = 1$ *this reduces to "independent events" or "no interference" with no restriction* (68), *and* (69) *reducing to* (60). *For* $\epsilon = 0$ *our model yields "complete interference" or "disjoint events" with restriction* (68) *becoming* (63) *and no restriction* (69). If we say that this model contains one parameter only, the idea is that the p_i are to be identified with the basic c.p. $c_{i,i+1}$. It might, however, seem adequate to consider ϵ and $p_1, \cdots, p_{m-1}$ as m available parameters which may be determined from the observations by some appropriate method.

(*iv*) *An* $(m - 1)$*-parametric model of partial interference.* Numerical data show (see particularly [4]) that interference is particularly marked i.e. $p_{ij} < p_ip_j$, if the corresponding p_i, p_j are very small, whereas for greater values of the p_i we have more nearly the pattern of independence. This is rather a striking fact, and seems to be well confirmed by observation. In these final pages a model will be studied which takes into account the circumstance that the amount of interference seems to depend on the magnitudes of the p_i. It contains $(m - 1)$ parameters, is therefore rather flexible, but nevertheless very simple.

Assume $m - 1 = n$ numbers ϵ_i where $0 \leqq \epsilon_i \leqq 1$ and form by means of n probabilities p_i:

$$\epsilon_i p_i = \bar{p}_i \qquad (0 \leqq \epsilon_i \leqq 1) \qquad (i = 1, 2, \cdots, n). \tag{70}$$

We may choose ϵ_i small if the corresponding p_i is small and larger if it is large; if the p's are all of the same order of magnitude the ϵ's need not differ much either. Then we simply define:

$$p_{ij} = \bar{p}_i\bar{p}_j, \qquad p_{ijk} = \bar{p}_i\bar{p}_j\bar{p}_k, \cdots, \qquad p_{12\ldots n} = \bar{p}_1\bar{p}_2 \cdots \bar{p}_n. \tag{71}$$

Let us investigate the consistency of this model. In analogy to (66) we form with $S_1 = \Sigma p_i$, $\bar{S}_1 = \Sigma \bar{p}_i$, $\bar{S}_2 = \Sigma \bar{p}_i \bar{p}_j$, etc.:

$$(72)\qquad \begin{aligned} P_0' &= 1 - S_1 + \bar{S}_2 - \bar{S}_3 + \cdots \\ &= (1 - \bar{S}_1 + \bar{S}_2 - \bar{S}_3 + \cdots) - \sum_{i=1}^{n} (1 - \epsilon_i)p_i = \bar{P}_0 - \sum_{i=1}^{n} (1 - \epsilon_i)p_i \end{aligned}$$

where P_0' and $\bar{P}_0$ are the probabilities for zero successes for the model under consideration and for independent events with basic probabilities $\bar{p}_i$ respectively; hence $\bar{P}_0 = \prod_{i=1}^{n} (1 - \epsilon_i p_i)$ and we get the condition:

$$(73)\qquad \prod_{i=1}^{n} (1 - \epsilon_i p_i) \geqq \sum_{i=1}^{n} (1 - \epsilon_i)p_i \quad \text{or:} \quad \sum_{i=1}^{n} p_i \leqq \sum_{i=1}^{n} \bar{p}_i + \prod_{i=1}^{n} (1 - \bar{p}_i).$$

If all $\epsilon_i = 1$ there is no restriction (73), while for $\epsilon_i = 0$ we find again (63). The consideration of P_1', P_2', $\cdots$ yields no new condition, because we get, denoting by $\bar{P}_i$ the probability of i successes for the independent events with basic probabilities $\bar{p}_i$:

$$\begin{aligned} P_1' &= S_1 - 2\bar{S}_2 + 3\bar{S}_3 - \cdots \pm n\bar{S}_n = \bar{S}_1 - 2\bar{S}_2 + \cdots \pm n\bar{S}_n + \sum_{i=1}^{n} p_i(1 - \epsilon_i) \\ &= \bar{P}_1 + \sum_{i=1}^{n} p_i(1 - \epsilon_i) \geqq 0 \quad \text{and:} \end{aligned}$$

$$P_i' = \bar{P}_i \geqq 0 \qquad (i \geqq 2).$$

As for the restrictions imposed by the linear theory we find:

$$(74)\qquad \begin{aligned} c_{i,j+1} - c_{ij} &= (1 - 2\bar{p}_i)(1 - 2\bar{p}_{i+1}) \cdots (1 - 2\bar{p}_{j-1})\bar{p}_j + p_j(1 - \epsilon_j) \\ c_{i,j+1} - c_{i+1,j+1} &= \bar{p}_i(1 - 2\bar{p}_{i+1}) \cdots (1 - 2\bar{p}_j) + p_i(1 - \epsilon_i). \end{aligned}$$

Thus the conditions of the linear theory are satisfied if

$$(75)\qquad \bar{p}_i \leqq \tfrac{1}{2} \quad \text{or} \quad p_i \leqq \frac{1}{2\epsilon_i}.$$

Hence summarizing: *On the basis of $m - 1$ probabilities p_i a consistent model of partial interference is obtained by means of* (70) *and* (71) *if the condition of consistency* (73) *and the conditions* (75) *are satisfied.*

It may be that the four simple models described in this section will seem too crude for the description of the complex mechanism of linkage. They could, of course, be combined and modified in various ways in order to serve at least as an approximation to the theoretical picture of reality we wish to construct. But, while these particular attempts may be inadequate, it seems to the author that the underlying principle is not wrong: that a mathematical theory of linkage must finally consist in statements on the l.d. (or the equivalent c.d.). The consideration of the c.p. is not sufficient for this purpose. The mathematical instrument for a theory of linkage seems to be the probability theory of the linkage distribution.

References

[1] F. Bernstein, *Variations- und Erblichkeitsstatistik*. Handbuch der Vererbungswissenschaft, Bd. 1, pp. 1-96.

[2] Kai Lai Chung, "On fundamental systems of probabilities of a finite number of events," *Ann. of Math. Stat.*, Vol. 14 (1943), pp. 234-37.

[3] H. Geiringer, (a) On the probability theory of arbitrarily linked events. *Ann. of Math. Stat.*, Vol. 9 (1938), pp. 260-271.

(b) "A note on the probability of arbitrary events," *Ann. of Math. Stat.*, Vol. 13 (1942) pp. 238-245.

[4] J. B. S. Haldane, (a) "The combination of linkage values and the calculation of distances between the loci of linked factors," *Jour. of Genetics*, Vol. 8 (1919), pp. 299-308.

(b) "Theoretical genetics of autopolyploids," *Jour. of Genetics*, Vol. 22 (1930), pp. 359-372.

[5] G. H. Hardy, "Mendelian proportions in a mixed population," *Science*, Vol. 28, (1908), p. 49-50.

[6] J. S. Huxley (editor), *The New Systematics*, Oxford 1940. (See articles by *S. Wright*, p. 161-183 and *H. J. Muller*, p. 158-268).

[7] H. S. Jennings, (a) "The numerical results of diverse systems of breeding with respect to two pairs of characters, etc.," *Genetics*, Vol. 12 (1917), pp. 97-154.

(b) "The numerical relations in the crossing over of the genes with a critical examination of the theory that the genes are arranged in a linear series," *Genetics*, Vol. 8 (1923), p. 393.

[8] K. v. Körösy, *Versuch einer Theorie der Genkoppelung*. (Bibliotheca Genetica), Leipzig 1929.

[9] K. Mather, *The Measurement of Linkage in Heredity*, London 1938.

[10] G. Mendel, "Versuche über Pflanzenhybriden." *Verh. des Naturforschd. Vereines in Brünn*, IV. Bd., *Abhandlungen* Brünn, 1866 pp. 3-47.

[11] T. H. Morgan, *The Theory of the Gene*. New Haven, 1928.

[12] K. Pearson, "On a generalized theory of alternative inheritance with special reference to Mendel's laws," *Phil. Trans. Royal Soc.* (A), Vol. 203 (1904), pp. 53-86.

[13] H. Rademacher, "Mathematische Theorie der Genkoppelung unter Berücksichtigung der Interferenz", 105. *Jahresber.* (1932), Schles. Ges. f. vaterländische Kultur. pp. 1-8.

[14] R. B. Robbins, (a) "Applications of mathematics to breeding problems, II." *Genetics*, Vol. 3 (1918), pp. 73-92.

(b) "Some applications of mathematics to breeding problems III." *Genetics*, Vol. 3 (1918), pp. 375-389.

[15] H. Tietze, "Über das Schicksal gemischter Populationen nach den Mendelschen Vererbungsgesetzen," *Zs. Angew. Math. u. Mech.*, Bd. 3, (1923), pp. 362-393.

[16] W. Weinberg, "Über Vererbungsgesetze beim Menschen." Zs. f. induktive Abstammungs- und Vererbungslehre, Vol. 1 (1909) p. 277-330.

[17] S. Wright, "Statistical genetics and evolution," *Bull. Amer. Math. Soc.*, Vol. 48 (1942), pp. 223-246.

THE COVARIANCE MATRIX OF RUNS UP AND DOWN

By H. Levene and J. Wolfowitz

Columbia University

1. Introduction. Let $a_1, \cdots, a_n$ be n unequal numbers and let the sequence $S = (h_1, h_2, \cdots, h_n)$ be any permutation of $a_1, \cdots, a_n$. S is to be considered a chance variable, and each of the $n!$ permutations of $a_1, \cdots, a_n$ is assigned the same probability. Consider the sequence R whose i^{th} element is the sign (+ or −) of $h_{i+1} - h_i$, $(i = 1, 2, \cdots, n-1)$. A sequence of p consecutive + signs not immediately preceded or followed by a + sign is called a run up of length p; a sequence of p consecutive − signs not immediately preceded or followed by a − sign is called a run down of length p. The term "run" will denote both runs up and runs down. The usage of the term "length" varies; most quality control literature attributes the length $p + 1$ to the runs which we say are of length p.

As an example of our usage, the sequence

$$S = 2\ 8\ 13\ 1\ 3\ 4\ 7$$

gives the sequence

$$R = + + - + + +,$$

which has a run up of length 2, followed by a run down of length 1, followed by a run up of length 3.

Runs up and down are widely used in quality control and have been applied to economic time series. The purpose of this paper is to obtain their variances and covariances and to correct some erroneous notions prevalent in the literature about their application.

2. Notation. If the sign (+ or −) of $(h_{i+1} - h_i)$ is the initial sign of a run defined as above, we call h_i the initial turning point (i. t. p.) of the run. Then h_1 is always an i. t. p., and we adopt the convention that h_n is never an i. t. p. We define new stochastic variables as follows:

$$x_i = \begin{cases} 1 & \text{if } h_i \text{ is an i. t. p.,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

$$x_{pi} = \begin{cases} 1 & \text{if } h_i \text{ is the i. t. p. of a run of length } p, \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

$$w_{pi} = \begin{cases} 1 & \text{if } h_i \text{ is the i. t. p. of a run of length } p \text{ or more,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

for $i = 1, 2, \cdots, n$. Also

$$r = \text{the number of runs in } R, \tag{2.4}$$

$$r_p = \text{the number of runs of length } p \text{ in } R, \tag{2.5}$$

$$r'_p = \text{the number of runs of length } p \text{ or more in } R. \tag{2.6}$$

Evidently $r = \sum_{i=1}^{n} x_i$, $r_p = \sum_{i=1}^{n} x_{pi}$, and $r'_p = \sum_{i=1}^{n} w_{pi}$.

If X and Y are stochastic variables, let $E(X)$ denote the mean of X, $\sigma(XY)$ denote the covariance of X and Y, and $\sigma^2(X)$ denote the variance of X, if they exist. By the distribution function of X we shall mean a function $f(x)$ such that $P\{X < x\} \equiv f(x)$, where the symbol $P\{\quad\}$ denotes the probability of the relation in the brackets.

3. Preliminary formulas. Let Y' be a stochastic variable with any continuous distribution function $f(y)$. Let $Y = (y_1, y_2 \cdots, y_n)$ be a sequence of n independent observations on Y'. Since $P\{y_i = y_j\} = 0$, $(i \neq j; i, j = 1, 2, \cdots, n)$, the distribution of runs up and down in Y is evidently the same as that in S. Now choose $f(y)$ to be

$$\begin{aligned} f(y) &= 0, & (y &\leq 0), \\ f(y) &= y, & (0 \leq y &\leq 1), \\ f(y) &= 1, & (y &\geq 1). \end{aligned}$$

Then

$$P\{y_{i-1} < y_i > y_{i+1}\} = \int_0^1 \left[\int_{y_{i+1}}^1 \left(\int_0^{y_i} dy_{i-1}\right) dy_i\right] dy_{i+1} = \frac{1}{3}.$$

By symmetry

$$E(x_i) = P\{y_{i-1} < y_i > y_{i+1}\} + P\{y_{i-1} > y_i < y_{i+1}\} = \tfrac{2}{3}, \qquad (i = 2, 3, \cdots, n-1).$$

Also $E(x_1) = 1$, and $E(x_n) = 0$.

It will be necessary hereafter to evaluate expressions of the types

$$U = \int_0^{y_{p+1}} \cdots \int_0^{y_2} \frac{(y_1)^k}{k!} dy_1 \cdots dy_p = \frac{(y_{p+1})^{k+p}}{(k+p)!}, \tag{3.1}$$

and

$$V = \int_{y_{p+1}}^1 \cdots \int_{y_2}^1 \frac{(y_1)^k}{k!} dy_1 \cdots dy_p. \tag{3.2}$$

From the fact that

$$\int_{y_{p+1}}^1 \cdots \int_{y_2}^1 dy_1 \cdots dy_p = \int_0^{v_{p+1}} \cdots \int_0^{v_2} dv_1 \cdots dv_p = \frac{(v_{p+1})^p}{p!}$$

where $v_j = (1 - y_j)$, $(j = 1, \cdots, p+1)$, it can easily be shown that

$$\begin{aligned} V &= \sum_{s=1}^{p} (-1)^{s+1} \frac{(v_{p+1})^{p-s}}{(k+s)!(p-s)!} + (-1)^p \frac{(y_{p+1})^{k+p}}{(k+p)!} \\ &= \sum_{r=0}^{k} (-1)^{k+r} \frac{(v_{p+1})^{p+k-r}}{(p+k-r)!\,r!}. \end{aligned} \tag{3.3}$$

We shall also need $\int_0^1 V\, dy_{p+1}$ and $\int_0^1 \int_0^{y_{p+2}} V\, dy_{p+1}\, dy_{p+2}$. Now

$$\int_0^1 V\, dy_{p+1} = \sum_{r=0}^{k} (-1)^{k+r} \frac{1}{(p+k-r+1)!\,r!}. \tag{3.4}$$

Making use of the relation,

$$\sum_{r=0}^{t} (-1)^r \frac{1}{(n-r)!\,r!} = (-1)^t \frac{1}{n(n-t-1)!\,t!}, \qquad (t < n),$$

we have

$$\int_0^1 V\,dy_{p+1} = \frac{1}{(p+k+1)p!\,k!}. \tag{3.5}$$

Similarly

$$\int_0^1 \int_0^{y_{p+2}} V\,dy_{p+1}\,dy_{p+2} = \frac{1}{(p+k+1)p!\,k!} - \frac{1}{(p+k+2)(p+1)!\,k!}. \tag{3.6}$$

4. Covariances of runs up and down. We first compute $E(r_p)$ and $E(r'_p)$. We define the symbol

$$P\{-, +^p, -\} = P\{y_{i-1} > y_i < y_{i+1} < \cdots < y_{i+p} > y_{i+p+1}\}.$$

The value of the right member is independent of i whenever it is defined (*i.e.* $i - 1 \geq 1, i + p + 1 \leq n$). Now

$$E(x_{pi}) = P\{-, +^p, -\} + P\{+, -^p, +\} = 2P\{-, +^p, -\}$$
$$= 2\int_0^1 \int_{y_{i+p+1}}^1 \int_0^{y_{i+p}} \cdots \int_0^{y_{i+1}} \int_{y_i}^1 dy_{i-1} \cdots dy_{i+p+1} = 2\frac{p^2+3p+1}{(p+3)!},$$
$$(i = 2, 3, \cdots, n-p-1).$$

$$E(x_{p1}) = 2P\{+^p, -\} \quad \text{and} \quad E(x_{p,n-p}) = 2P\{-, +^p\}.$$

By symmetry $E(x_{p1}) = E(x_{p,n-p})$, the common value being $2\frac{p+1}{(p+2)!}$. Also $E(x_{pi}) = 0, (i > n - p)$.
Thus

$$E(r_p) = E\left(\sum_{i=1}^{n} x_{pi}\right) = 2E(x_{p1}) + (n-p-2)E(x_{pi})$$
$$= 2n\frac{p^2+3p+1}{(p+3)!} - 2\frac{p^3+3p^2-p-4}{(p+3)!}, \qquad (p \leq n-2). \tag{4.1}$$

Besson [1], Kermack and McKendrick [5], and Wallis and Moore [6] gave the exact value, although Besson proved it only for special cases. R. A. Fisher [3] gave $\lim_{n\to\infty} \frac{E(r_p)}{E(r)}$.

It is clear that $E(w_{pi}) = E(x_{p,n-p})$, $(i = 2, \cdots, n-p)$, while $E(w_{p1}) = 2P\{+^p\} = 2/(p-1)!$. We then have

$$E(r'_p) = 2n\frac{p+1}{(p+2)!} - 2\frac{p^2+p-1}{(p+2)!}, \qquad (p \leq n-1). \tag{4.2}$$

Setting $p = 1$ we have

$$E(r_1') = E(r) = \tfrac{1}{3}(2n - 1). \tag{4.3}$$

Formula (4.3) was given by Bienaymé [2].

We now obtain $\sigma(r_p r_q)$. Let $(x_{pi} - E(x_{pi})) = z_{pi}$. Then

$$\begin{aligned} \sigma(r_p r_q) &= E\left\{\left[\sum_{i=1}^{n} z_{pi}\right]\left[\sum_{j=1}^{n} z_{qj}\right]\right\} \\ &= \sum_i E(z_{pi} z_{qi}) + \sum_{i<j}\sum E(z_{qi} z_{pj}) + \sum_{i<j}\sum E(z_{pi} z_{qj}). \end{aligned} \tag{4.4}$$

For $j \geq i + q + 3$, x_{qi} and x_{pj} are independent and hence $E(z_{qi}z_{pj}) = 0$. Omitting zero terms from (4.4) we have

$$\begin{aligned} \sigma(r_p r_q) = \{\sum_i E(x_{pi} x_{qi}) &+ \sum_{i<j<i+q+3}\sum E(x_{qi} x_{pj}) + \sum_{i<j<i+p+3}\sum E(x_{pi} x_{qj}) \\ &- [\sum_i E(x_{pi})E(x_{qi}) + \sum_{i<j<i+q+3}\sum E(x_{qi})E(x_{pj}) \\ &+ \sum_{i<j<i+p+3}\sum E(x_{pi})E(x_{qj})]\}. \end{aligned} \tag{4.5}$$

Since $x_{pi}x_{qi} = \delta_{pq}(x_{pi})^2 = \delta_{pq}x_{pi}$, we have for the first term of the right member of (4.5)

$$\sum_{i=1}^{n} E(x_{pi} x_{qi}) = \delta_{pq} E(r_p), \tag{4.6}$$

where the Kronecker delta $\delta_{pq} = \begin{cases} 1, \text{ if } p = q, \\ 0, \text{ otherwise.} \end{cases}$

Since $x_{qi}x_{pj} = 0$ for $i < j < i + q$, the second term in the right member of (4.5) is

$$\begin{aligned} &\sum_{i=1}^{n-p-q} E(x_{qi} x_{p,i+q}) + \sum_{i=1}^{n-p-q-1} E(x_{qi} x_{p,i+q+1}) + \sum_{i=1}^{n-p-q-2} E(x_{qi} x_{p,i+q+2}) \\ &\quad = \{(n - p - q - 2)E(x_{qi} x_{p,i+q}) + (n - p - q - 3)E(x_{qi} x_{p,i+q+1}) \\ &\quad + (n - p - q - 4)E(x_{qi} x_{p,i+q+2}) \\ &\quad + E(x_{q1} x_{p,q+1}) + E(x_{q1} x_{p,q+2}) + E(x_{q1} x_{p,q+3}) \\ &\quad + E(x_{q,n-q-p} x_{p,n-p}) + E(x_{q,n-q-p-1} x_{p,n-p}) + E(x_{q,n-q-p-2} x_{p,n-p})\}. \end{aligned} \tag{4.7}$$

Now $E(x_{qi}x_{p,i+q}) = 2P\{-, +^q, -^p, +\} = 2A$, where

$$A = \int_0^1 \int_0^{y_{i+p+q+1}} \int_{y_{i+p+q}}^1 \cdots \int_{y_{i+q+1}}^1 \left[\int_0^{y_{i+q}} \cdots \int_0^{y_{i+1}} \int_{y_i}^1 dy_{i-1} \cdots dy_{i+q-1}\right] \cdot dy_{i+q} \cdots dy_{i+p+q+1} .$$

The expression within the square brackets is easily evaluated, and applying (3.6) to the result, we have

$$\begin{aligned} A = {} & \frac{1}{(p+q+1)p!q!} - \frac{1}{(p+q+2)(p+1)!q!} \\ & - \frac{1}{(p+q+2)p!(q+1)!} + \frac{1}{(p+q+3)(p+1)!(q+1)!}. \end{aligned}$$

Similarly, $E(x_{qi}x_{p,i+q+1}) = 2P\{-, +^q, -, +^p, -\}$, and $E(x_{qi}x_{p,i+q+2}) = 2P\{-, +^q, -, -, +^p, -\} + 2P\{-, +^q, -, +, -^p, +\}$. The other terms in the right member of (4.7) are obtained in like manner. The right member of (4.7) is symmetric in p and q; hence the second and third terms of the right member of (4.5) are equal.

We now consider the remaining terms in the right member of (4.5) for $p > q$; the result obtained also holds for $p \leq q$. We write them as

$$\begin{aligned} & -\Bigg\{\sum_{i=p+3}^{n-q} E(x_{qi})E(x_{p,i-(p+2)}) + \sum_{i=p+2}^{n-q} E(x_{qi})E(x_{p,i-(p+1)}) \\ & \quad + \cdots + \sum_{i=p-q+1}^{n-q} E(x_{qi})E(x_{p,i-(p-q)}) + \sum_{i=p-q}^{n-q-1} E(x_{qi})E(x_{p,i-(p-q-1)}) \\ & \quad + \cdots + \sum_{i=1}^{n-p} E(x_{qi})E(x_{pi}) + \sum_{i=1}^{n-p-1} E(x_{qi})E(x_{p,i+1}) \\ & \quad + \cdots + \sum_{i=1}^{n-p-(q+2)} E(x_{qi})E(x_{p,i+(q+2)})\Bigg\}. \end{aligned} \tag{4.8}$$

The $(p + q + 5)$ sums in (4.8) comprise in all $\left\{(n - p)(p + q + 5) - 2\sum_{k=1}^{q+2} k\right\}$ terms. Remembering that $E(x_{p,n-p}) = E(x_{p1})$, (4.8) becomes

$$\begin{aligned} & -\{[n(p+q+5) - (p^2 + pq + q^2 + 7p + 7q + 16)]E(x_{qi})E(x_{pj}) \\ & \quad + [2p+4]E(x_{qi})E(x_{p1}) + [2q+4]E(x_{q1})E(x_{pj}) + 2E(x_{q1})E(x_{p1})\}. \end{aligned} \tag{4.9}$$

Adding the right member of (4.6), twice the right member of (4.7), and (4.9), we have

$$\begin{aligned} & \sigma(r_p r_q) = \\ & 2n\Bigg\{-2\,\frac{\begin{array}{c} p^2(p+q+6)(q^2+3q+1) \\ + p(3q^3+20q^2+40q+19) + (q^3+9q^2+29q+26) \end{array}}{(q+3)!(p+3)!} \\ & \quad + 2\,\frac{-p+q+1}{(p+q+3)(q+2)!(p+1)!} - 2\,\frac{1}{(p+q+5)(q+3)!(p+1)!} \\ & \quad - 2\,\frac{(p+q)^3 + 9(p+q)^2 + 23(p+q) + 14}{(p+q+5)!} \\ & \quad + 2\,\frac{1}{(p+q+1)q!p!} + \delta_{pq}\,\frac{p^2+3p+1}{(p+3)!}\Bigg\} \end{aligned}$$

$$\begin{aligned}
&+2\left\{\frac{2}{(q+3)!(p+3)!}\left[\begin{aligned}&p^4(q^2+3q+1)+p^3(q^3+9q^2+19q+6)\\ &\quad+p^2(q^4+9q^3+28q^2+35q+11)\\ &\quad+p(3q^4+20q^3+40q^2+29q+10)\\ &\quad+(q^4+9q^3+27q^2+32q+10)\end{aligned}\right]\right.\\
&+2\,\frac{(p+q+2)(p-q-1)}{(p+q+3)(q+2)!(p+1)!}+2\,\frac{p+q+4}{(p+q+5)(q+3)!(p+1)!}\\
&+2\,\frac{(p+q)^4+10(p+q)^3+29(p+q)^2+16(p+q)-19}{(p+q+5)!}\\
&\left.-2\,\frac{p+q}{(p+q+1)q!p!}-\delta_{pq}\,\frac{p^3+3p^2-p-4}{(p+3)!}\right\},
\end{aligned}\tag{4.10}$$

where δ_{pq} is the Kronecker delta. Formula (4.10) is valid for $p+q \leq n-4$. It is symmetric in p and q. Setting $p = q$ we obtain

$$\begin{aligned}
\sigma^2(r_p) = 2n&\left\{-2\,\frac{2p^5+15p^4+41p^3+55p^2+48p+26}{(p+3)!(p+3)!}\right.\\
&+2\,\frac{2p^2+9p+12}{(2p+3)(2p+5)(p+3)!(p+1)!}-4\,\frac{4p^3+18p^2+23p+7}{(2p+5)!}\\
&\left.+2\,\frac{1}{(2p+1)p!p!}+\frac{p^2+3p+1}{(p+3)!}\right\}\\
&+2\left\{2\,\frac{3p^6+24p^5+69p^4+90p^3+67p^2+42p+10}{(p+3)!(p+3)!}\right.\\
&-4\,\frac{2p^3+11p^2+19p+9}{(2p+3)(2p+5)(p+3)!(p+1)!}\\
&+2\,\frac{16p^4+80p^3+116p^2+32p-19}{(2p+5)!}\\
&\left.-4\,\frac{p}{(2p+1)p!p!}-\frac{p^3+3p^2-p-4}{(p+3)!}\right\}
\end{aligned}\tag{4.11}$$

We next evaluate $\sigma(r'_p r'_q)$. Since w_{qi} and w_{pj} are independent for $j \geq i+q+2$, we have, corresponding to (4.5),

$$\begin{aligned}
\sigma(r'_p r'_q) = \sum_i E(w_{pi}w_{qi}) &+ \sum_{i<j<}\sum_{i+q+2} E(w_{qi}w_{pj}) + \sum_{i<j<}\sum_{i+p+2} E(w_{pi}w_{qj})\\
&- [\sum_i E(w_{pi})E(w_{qi}) + \sum_{i<j<}\sum_{i+q+2} E(w_{qi})E(w_{pj})\\
&+ \sum_{i<j<}\sum_{i+p+2} E(w_{pi})E(w_{qj})].
\end{aligned}\tag{4.12}$$

Let $G = \text{Max}\,(p, q)$. Then $w_{pi}w_{qi} \equiv w_{Gi}$ and we have for the first term of the right member of (4.12)

$$\sum_{i=1}^{n} E(w_{pi}w_{qi}) = E(r'_G). \tag{4.13}$$

The second term in the right member of (4.12) may be written

$$\begin{aligned}(n - p - q - 1)E(w_{qi}w_{p,i+q}) + (n - p - q - 2)E(w_{qi}w_{p,i+q+1})\\ + E(w_{q1}w_{p,q+1}) + E(w_{q1}w_{p,q+2}).\end{aligned} \tag{4.14}$$

Now $E(w_{qi}w_{p,i+q}) = 2P\{-, +^q, -^p\}$, $E(w_{qi}w_{p,i+q+1}) = 2P\{-, +^q, -, +^p\} + 2P\{-, +^{q+1}, -^p\}$, and the other terms are obtained similarly. The third term in the right member of (4.12) will be equal to (4.14) with p and q interchanged.

The remaining terms in the right member of (4.12) become

$$\begin{aligned}&-\{[n(p + q + 3) - (p^2 + pq + q^2 + 4p + 4q + 5)]E(w_{qi})E(w_{pj})\\ &+ [p + 1]E(w_{qi})E(w_{p1}) + [q + 1]E(w_{q1})E(w_{pj}) + E(w_{q1})E(w_{p1})\}.\end{aligned} \tag{4.15}$$

We can now write the formula for $\sigma(r'_p r'_q)$, valid for $p + q \leq n - 2$,

$$\begin{aligned}\sigma(r'_p r'_q) = 2n \Bigg\{&- \frac{p^2(2q + 2) + p(2q^2 + 8q + 5) + (2q + 1)(q + 2)}{(q + 2)!(p + 2)!}\\ &+ \frac{2}{(p + q + 1)q!p!} - \frac{(q + 1)(q + 2) + (p + 1)(p + 2)}{(p + q + 3)(q + 2)!(p + 2)!}\\ &- 2\,\frac{p + q + 2}{(p + q + 3)!} + \frac{(G + 1)}{(G + 2)!}\Bigg\}\\ + 2\Bigg\{&\frac{p^3(2q + 2) + p^2(2q^2 + 8q + 5) + p(2q^3 + 8q^2 + 6q - 2) + (2q^3 + 5q^2 - 2q - 6)}{(q + 2)!(p + 2)!}\\ &- 2\,\frac{p + q}{(p + q + 1)q!p!} + \frac{(p + q + 2)[(p + 1)(p + 2) + (q + 1)(q + 2)]}{(p + q + 3)(q + 2)!(p + 2)!}\\ &+ 2\,\frac{(p + q)^2 + 3(p + q) + 1}{(p + q + 3)!} - \frac{G^2 + G - 1}{(G + 2)!}\Bigg\},\end{aligned} \tag{4.16}$$

where $G = \text{Max}\,(p, q)$. Setting $p = q$ we obtain

$$\begin{aligned}\sigma^2(r'_p) = 2n \Bigg\{&- 2\,\frac{(p + 1)(2p^2 + 4p + 1)}{(p + 2)!(p + 2)!} + 2\,\frac{1}{(2p + 1)p!p!}\\ &- 2\,\frac{1}{(2p + 3)(p + 2)!p!} - 4\,\frac{p + 1}{(2p + 3)!} + \frac{p + 1}{(p+2)!}\Bigg\}\\ &+ 2\Bigg\{2\,\frac{(p + 1)^2(3p^2 + 4p - 3)}{(p + 2)!(p + 2)!} - 4\,\frac{p}{(2p + 1)p!p!}\\ &+ 4\,\frac{p + 1}{(2p + 3)(p + 2)!p!} + 2\,\frac{4p^2 + 6p + 1}{(2p + 3)!} - \frac{p^2 + p - 1}{(p + 2)!}\Bigg\}.\end{aligned} \tag{4.17}$$

Setting $p = 1$, we have

$$\sigma^2(r_1') = \sigma^2(r) = \frac{16n - 29}{90}. \tag{4.18}$$

The value of $\sigma^2(r)$ was given by Bienaymé [2].

Finally, we evaluate

$$\begin{aligned} \sigma(r_p r_q') = {} & \sum_i E(x_{pi} w_{qi}) + \sum_{i<j<i+q+2}\sum E(w_{qi} x_{pj}) + \sum_{i<j<i+p+3}\sum E(x_{pi} w_{qj}) \\ & - [\sum_i E(x_{pi})E(w_{qi}) + \sum_{i<j<i+q+2}\sum E(w_{qi})E(x_{pj}) \\ & + \sum_{i<j<i+p+3}\sum E(x_{pi})E(w_{qj})]. \end{aligned} \tag{4.19}$$

Let the symbol $\eta_{ij} = \begin{Bmatrix} 1, i \geq j \\ 0, i < j \end{Bmatrix}$. Then $x_{pi} w_{qi} \equiv \eta_{pq} x_{pi}$, and

$$\sum_{i=1}^{n} E(x_{pi} w_{qi}) = \eta_{pq}[E(r_p)]. \tag{4.20}$$

The remaining terms of (4.19) introduce no new difficulties, and for $p + q \leq n - 3$ we obtain

$$\begin{aligned} \sigma(r_p r_q') = 2n \Bigg\{ & - \frac{p^3(2q+2) + p^2(2q^2+13q+12) + p(6q^2+22q+23) + (2q^2+6q+15)}{(p+3)!\,(q+2)!} \\ & + \frac{2}{(p+q+1)p!\,q!} + \frac{p-q}{(p+q+2)(p+1)!(q+1)!} \\ & + \frac{(p-q+1)(q+2)}{(p+q+3)(p+2)!(q+2)!} + \frac{(p+2)(p+3)+(q+1)(q+2)}{(p+q+4)(p+3)!(q+2)!} \\ & - 2\,\frac{(p+q)^2 + 5(p+q) + 5}{(p+q+4)!} + \eta_{pq}\,\frac{p^2+3p+1}{(p+3)!} \Bigg\} \\ & + 2\Bigg\{ \frac{1}{(p+3)!(q+2)!} \begin{bmatrix} p^4(2q+2) + p^3(2q^2+13q+12) \\ + p^2(2q^3+13q^2+26q+24) \\ + p(6q^3+22q^2+19q+27) \\ + (2q^3+6q^2+10q+25) \end{bmatrix} \\ & - 2\,\frac{p+q}{(p+q+1)p!\,q!} - \frac{p^2+2p-q^2+2}{(p+q+2)(p+1)!(q+1)!} \\ & - \frac{(p+2)[(p+2)(q+3)-1] - q(q+1)(q+2)}{(p+q+3)(p+2)!(q+2)!} \\ & - \frac{(p+q+3)[(p+2)(p+3)+(q+1)(q+2)]}{(p+q+4)(p+3)!(q+2)!} \\ & + 2\,\frac{(p+q)^3 + 6(p+q)^2 + 8(p+q) - 1}{(p+q+4)!} - \eta_{pq}\,\frac{p^3+3p^2-p-4}{(p+3)!} \Bigg\}, \end{aligned} \tag{4.21}$$

where η_{pq} is defined as in (4.20).

5. The use of runs up and down in tests of significance. Certain misconceptions about the application of runs up and down have appeared in the literature, and it is the purpose of this section to clarify them.

Since $E(r_p)$, $\sigma^2(r_p)$, $E(r)$ and $\sigma^2(r)$ are all of the order n, it follows that r_p/r converges stochastically to

$$\lambda_p = \lim_{n\to\infty} \frac{E(r_p)}{E(r)} .$$

Let

$$\lambda'_p = \lim_{n\to\infty} \frac{E(r'_p)}{E(r)} .$$

From (4.1) and (4.2) we have

$$\lambda_1 = \frac{5}{8} = .6250$$

$$\lambda_2 = \frac{11}{40} = .2750$$

$$\lambda_3 = \frac{19}{240} = .07917$$

$$\lambda_4 = \frac{29}{1680} = .01726$$

$$\lambda'_5 = \frac{1}{280} = .00357$$

Let

$$\lambda_{pn} = \frac{E(r_p)}{E(r)} .$$

Some writers say that λ_{pn} or λ_p is "the probability of a run of length p." If the stochastic process consists in obtaining a sequence from among the $n!$ sequences S, each of which has the probability $(n!)^{-1}$, then the phrase "the probability of *a* run of length p" has no meaning. One can speak of the probability of at least one run of length p (i.e., that $r_p > 0$), of the probability of no run of length p ($r_p = 0$), of the probability that the first or fifth run (if there are five runs) in the sequence S be of length p, etc. It is possible to give *different* stochastic processes in which "the probability of a run of length p" will have meaning and be λ_{pn}, or λ_p. Consider, for example, the totality of all the *runs* in the $n!$ sequences S. There are $n!E(r)$ of them, and among these there are $n!E(r_p)$ runs of length p. Now let the stochastic process consist in drawing a *run* from the totality of all these runs, each of which is to have the same probability, which is therefore $[n!E(r)]^{-1}$. Then the probability of drawing a run of length p is λ_{pn}. It is difficult to see how this stochastic process can have relevance to most of the problems of quality control and economic time series where runs up and down are now used.

Some writers on quality control and economic time series recommend that statistical control or randomness be tested by use of $d_1, \cdots, d_{p-1}, d'_p$, where

$$\begin{aligned} d_i &= r_i - E(r_i), \qquad (i = 1, 2, \cdots, (p-1)), \\ d'_p &= r'_p - E(r'_p). \end{aligned}$$

The availability of the covariance matrix M of $d_1, \cdots, d_{p-1}, d'_p$, which we have obtained in this paper, will assist in the construction of such tests. Also of help will be a result recently announced by one of us [7], the early publication of which is expected. This result states that in the limit with n the joint probability density function of $d_1, \cdots, d_{p-1}, d'_p$, is $Ke^{-\frac{1}{2}Q}$, where K is a constant and Q is a quadratic form in $d_1, \cdots, d_{p-1}, d'_p$ whose matrix is the inverse of the matrix M. It follows immediately that Q has in the limit the χ^2 distribution with p degrees of freedom.

We wish now to make a few remarks about the tests of significance, based on runs up and down, which are used by some contemporary writers. A description of their method can perhaps be best given by an example. With $n = 100$ and $p = 5$, say, suppose the observed values are:

Observed Values

$$\begin{array}{rl} r_1 &= 30 \\ r_2 &= 10 \\ r_3 &= \ \ 4 \\ r_4 &= \ \ 3 \\ r'_5 &= \ \ 3 \\ \hline \text{Total, } r &= 50 \end{array}$$

These writers then say that the expected values are:

Expected Values according to some writers

$$\begin{array}{lllr} E(r_1) = r\lambda_1 &= 50\ (.6250) &= & 31.25 \\ E(r_2) = r\lambda_2 &= 50\ (.2750) &= & 13.75 \\ E(r_3) = r\lambda_3 &= 50\ (.07917) &= & 3.96 \\ E(r_4) = r\lambda_4 &= 50\ (.01726) &= & 0.86 \\ E(r'_5) = r\lambda'_5 &= 50\ (.00357) &= & 0.18 \\ \hline & & & 50.00 \end{array}$$

The correct expected values are given by (4.1) and (4.2) and are:

Correct Expected Values

$$\begin{array}{lr} E(r_1) = & 41.75 \\ E(r_2) = & 18.10 \\ E(r_3) = & 5.15 \\ E(r_4) = & 1.11 \\ E(r'_5) = & 0.22 \\ \hline & 66.33 \end{array}$$

It should be noted that:

(a) A consequence of the incorrect method of obtaining "expected values" is that, since

$$E(r) = E(r_1) + E(r_2) + E(r_3) + E(r_4) + E(r_5'),$$

it implies that the *expected* number of runs of all lengths is equal to the *observed* number! This is obviously erroneous. In fact it follows from (4.18) and the results announced in [7] that $r - E(r)$ is in the probability sense of order $\sqrt{n}$.

(b) By using the incorrect expected values for comparison with the observed values one loses the valuable information furnished by $r - E(r)$. If this is large (in terms of its standard deviation) it is plausible to question whether statistical control or randomness exists.

6. Summary. Let $S = (h_1, \cdots, h_n)$ be a random permutation of the n unequal numbers $a_1, \cdots, a_n$, and let R be the sequence of signs ($+$ or $-$) of the differences $h_{i+1} - h_i$ $(i = 1, \cdots, n - 1)$. It is assumed that each of the $n!$ sequences S is equally probable. A sequence of p successive $+$ $(-)$ signs not immediately preceded or followed by a $+$ $(-)$ sign is called a run up (down) of length p. Let r_p and r_p' be the number of runs up and down in R of lengths p and p or more respectively. In this paper the exact values of $\sigma(r_p r_q)$, (see formula (4.10)); $\sigma^2(r_p)$, (formula (4.11)); $\sigma(r_p' r_q')$, (formula (4.16)); $\sigma^2(r_p')$, (formula (4.17)); and $\sigma(r_p r_q')$, (formula (4.21)) are derived. A few numerical values are:

$$\sigma^2(r_1) = \frac{305n - 347}{720}, \qquad \sigma^2(r_2) = \frac{51106n - 73859}{453600},$$

$$\sigma^2(r_1') = \frac{16n - 29}{90}, \qquad \sigma^2(r_2') = \frac{57n - 43}{720}, \qquad \sigma^2(r_3') = \frac{21496n - 51269}{453600},$$

$$\sigma(r_1 r_2) = -\frac{19n + 11}{210}, \qquad \sigma(r_1' r_2') = -\frac{5n - 3}{60}, \qquad \sigma(r_1' r_3') =$$

$$-\frac{41n - 99}{630}, \qquad \sigma(r_2 r_1') = -\frac{23n + 135}{1260}, \qquad \sigma(r_1 r_2') = -\frac{117n - 79}{720},$$

$$\sigma(r_1 r_3') = -\frac{363n - 817}{5040}, \qquad \text{and} \quad \sigma(r_2 r_3') = -\frac{18346n - 49019}{453600}.$$

The values of $E(r_p)$, (formula (4.1)); and $E(r_p')$, (formula (4.2)) are also given. Certain misconceptions about the applications of runs up and down are discussed.

REFERENCES

[1] L. Besson, (transl. and abr. by E. W. Woolard), "On the comparison of meteorological data with chance results," *Monthly Weather Review*, (U. S. Weather Bureau), Vol. 48 (1920), pp. 89–94.

[2] J. Bienaymé, "Sur une question de probabilités," *Bull. Soc. Math. France*, Vol. 2 (1874), pp. 153–154.

[3] R. A. Fisher, "On the random sequence," *Quarterly Jour. Roy. Meteorological Soc.* Vol. 52 (1926), p. 250.

[4] W. O. Kermack and A. G. McKendrick, "Tests for randomness in a series of observations," *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp. 228–240.

[5] W. O. Kermack and A. G. McKendrick, "Some distributions associated with a randomly arranged set of numbers," *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp. 332–376.

[6] W. A. Wallis and G. H. Moore, *A significance test for time series*, Technical Paper 1, Nat. Bureau of Econ. Res., New York, 1941.

[7] J. Wolfowitz, "Asymptotic distributions of ascending and descending runs," [abstract] *Bull. Amer. Math. Soc.*, Vol. 49 (1943), pp. 539–540.

ON THE MEASURE OF A RANDOM SET

By H. E. Robbins

Post Graduate School, U. S. Naval Academy

1. Introduction. The following is perhaps the simplest non-trivial example of the type of problem to be considered in this paper. On the real number axis let N points x_i $(i = 1, 2, \cdots, N)$ be chosen independently and by the same random process, so that the probability that x_i shall lie to the left of any point x is a given function of x,

$$\sigma(x) = \Pr\,(x_i < x). \tag{1}$$

With the points x_i as centers, N unit intervals are drawn. Let X denote the set-theoretical sum of the N intervals, and let $\mu(X)$ denote the linear measure of X. Then $\mu(X)$ will be a chance variable whose values may range from 1 to N, and whose probability distribution is completely determined by $\sigma(x)$. Let $\tau(u)$ denote the probability that $\mu(X)$ be less than u. Then by definition, the expected value of $\mu(X)$ is

$$E(\mu(X)) = \int_1^N u\,d\tau(u), \tag{2}$$

where

$$\tau(u) = \Pr\,(\mu(X) < u). \tag{3}$$

The problem is to transform the expression for $E(\mu(X))$ so that its value may be computed in terms of the given function $\sigma(x)$.

In order to do this, we observe that, since the x_i are independent,

$$\tau(u) = \int \underset{C(u)}{\cdots} \int d\sigma(x_1) \cdots d\sigma(x_N), \tag{4}$$

where the domain of integration $C(u)$ consists of all points $(x_1, \cdots, x_N)$ in Euclidean N-dimensional space such that the linear measure of the set-theoretical sum of N unit intervals with centers at the points x_i is less than u. Here, however, a difficulty arises. Due to the possible overlapping of the intervals, the geometrical description of the domain $C(u)$ is such as to make the explicit evaluation of the integral (4) a complicated matter.

The difficulty is even more serious in the analogous problem where instead of N unit intervals on the line we have N unit circles in the plane, with a given probability distribution for their centers (x_i, y_i). Again we seek the expected value of the measure of the set-theoretical sum of the N circles. The corresponding domain $C(u)$ in $2N$-dimensional space will now be very complicated.

It is the object of this paper to show how, in such cases as these, the expected value of $\mu(X)$ may be found without first finding the distribution function $\tau(u)$.

In fact, the theorem to be stated in (15) will in many important cases yield a comparatively simple formula for $E(\mu(X))$.

2. Expected value of $\mu(X)$. In order to state the problem in full generality, let us suppose that X is a random Lebesgue measurable subset of Euclidean n dimensional space E_n. By this we shall mean that in the space T of all possible values of X there is defined a probability measure $\rho(X)$ so that for every ρ-measurable subset S of T, the probability that X shall belong to S is given by the Lebesgue-Stieltjes integral

$$\Pr(X \epsilon S) = \int_T C_S(X)\, d\rho(X), \tag{5}$$

where the integrand is the characteristic function of S,

$$C_S(X) = \begin{cases} 1 & \text{for} \quad X \epsilon S \\ 0 & \text{for} \quad X \notin S. \end{cases} \tag{6}$$

In practice, the set X will be a function of a finite number of real parameters (e.g., the coordinates of the centers of the intervals or circles considered in the Introduction), $X = X(\alpha_1, \cdots, \alpha_r) = X(\alpha)$. There will be given a probability measure $\nu(\alpha)$ in the parameter space E_r, so that α will be a vector random variable in the ordinary sense. If A is any ν-measurable subset of E_r, then by definition,

$$\Pr(\alpha \epsilon A) = \int_{E_r} C_A(\alpha)\, d\nu(\alpha). \tag{7}$$

Now for the set S' consisting of all X such that $X = X(\alpha)$ for α in A, we define $\rho(S') = \nu(A)$. Thus a ρ-measure is defined in the space T of X, which is the general situation considered in the preceding paragraph.

Returning to the general case described in the first paragraph of this section, we shall now prove the main theorem of this paper. To this end we define, for every point x of E_n and every set X of T, the function

$$g(x, X) = \begin{cases} 1 & \text{for} \quad x \epsilon X \\ 0 & \text{for} \quad x \notin X. \end{cases} \tag{8}$$

Moreover, for every x in E_n we let $S(x)$ denote the set of all X in T which contain x. Then for every x in E_n we have from (6),

$$g(x, X) = C_{S(x)}(X). \tag{9}$$

Let us denote the Lebesgue measure in E_n of the set X by $\mu(X)$. Assuming that the function $g(x, X)$ is a $\mu\rho$-measurable function of the pair (x, X) in the product space[1] of E_n with T, it follows from Fubini's theorem[1] that

$$\int_{E_n \times T} g(x, X)\, d\mu\rho(x, X) = \int_{E_n} \int_T g(x, X)\, d\rho(X)\, d\mu(x). \tag{10}$$

[1] See S. Saks, *Theory of the Integral*, G. E. Stechert, N. Y., 1937, pp. 86, 87.

From (5) and (9) it follows that

$$\int_T g(x, X)\, d\rho(X) = \Pr\,(X \epsilon S(x)) = \Pr\,(x \epsilon X). \tag{11}$$

Again by Fubini's theorem we have

$$\int_{E_n \times T} g(x, X)\, d\mu\rho(x, X) = \int_T \int_{E_n} g(x, X)\, d\mu(x)\, d\rho(X). \tag{12}$$

But from (8),

$$\int_{E_n} g(x, X)\, d\mu(x) = \int_X d\mu(x) = \mu(X). \tag{13}$$

Now from (10), (11), (12), and (13) we have

$$\int_{E_n} \Pr\,(x \epsilon X)\, d\mu(x) = \int_T \mu(X)\, d\rho(X). \tag{14}$$

But the latter integral is equal to $E(\mu(X))$. Hence we have the relation

$$E(\mu(X)) = \int_{E_n} \Pr\,(x \epsilon X)\, d\mu(x). \tag{15}$$

This is our fundamental result. We may state it as a

THEOREM: *Let X be a random Lebesgue measurable subset of E_n, with measure $\mu(X)$. For any point x of E_n let $p(x) = \Pr\,(x \epsilon X)$. Then, assuming that the function $g(x, X)$ defined by* (8) *is a measurable function of the pair (x, X), the expected value of the measure of X will be given by the Lebesgue integral of the function $p(x)$ over E_n.*

3. Higher moments of $\mu(X)$. We may generalize the result (15) to obtain similar expressions for the higher moments of $\mu(X)$. For the second moment we have the expression

$$E(\mu^2(X)) = \int_T \mu^2(X)\, d\rho(X). \tag{16}$$

Now from (13),

$$\begin{aligned} \mu^2(X) = \mu(X)\cdot\mu(X) &= \int_{E_n} g(x, X)\, d\mu(x) \cdot \int_{E_n} g(y, X)\, d\mu(y) \\ &= \int_{E_n} \int_{E_n} g(x, X)\cdot g(y, X)\, d\mu(x)\, d\mu(y). \end{aligned} \tag{17}$$

Let

$$g(x, y, X) = g(x, X)\cdot g(y, X) \begin{array}{l} = 1 \text{ if } X \text{ contains both } x \text{ and } y \\ = 0 \text{ otherwise.} \end{array} \tag{18}$$

Then from (16), (17), and (18), we have as before by Fubini's theorem,

$$
\begin{aligned}
E(\mu^2(X)) &= \int_T \int_{E_n} \int_{E_n} g(x, y, X)\, d\mu(x)\, d\mu(y)\, d\rho(X) \\
&= \int_{E_n} \int_{E_n} \int_T g(x, y, X)\, d\rho(X)\, d\mu(x)\, d\mu(y).
\end{aligned} \tag{19}
$$

But from (5) and (18) it follows that

$$\int_T g(x, y, X)\, d\rho(X) = \Pr\ (x \epsilon X \text{ and } y \epsilon X). \tag{20}$$

The latter probability may be denoted by $p(x, y)$. This function will be defined over the Cartesian product, E_{2n}, of E_n with itself. Let $\mu(x, y)$ denote Lebesgue measure in E_{2n}. Then from (19) we have

$$E(\mu^2(X)) = \int_{E_{2n}} p(x, y)\, d\mu(x, y), \tag{21}$$

where

$$p(x, y) = \Pr\ (x \epsilon X \text{ and } y \epsilon X). \tag{22}$$

The formula for the mth moment of $\mu(X)$ will clearly be

$$\text{Exp}\ (\mu^m(X)) = \int_{E_{mn}} p(x_1, x_2, \cdots, x_m)\, d\mu(x_1, x_2, \cdots, x_m), \tag{23}$$

where $\mu(x_1, x_2, \cdots, x_m)$ denotes Lebesgue measure in E_{mn} and where

$$p(x_1, x_2, \cdots, x_m) = \Pr\ (x_1 \epsilon X \text{ and } x_2 \epsilon X \cdots \text{and } x_m \epsilon X). \tag{24}$$

In the next section we shall apply formulas (15) and (21) to a specific problem.

4. Let a, p, B be given positive numbers such that $(B + a)p \leq a$ and $a \leq B$. We shall define the random linear point set X as follows. N intervals, each of length a, are chosen independently on the number axis. The probability density function for the center of the ith interval will be assumed to be constant and equal to p/a in the interval $-a/2 \leq x \leq B + (a/2)$; it may be arbitrary outside this interval. The set X is now defined as the intersection of the fixed interval I: $0 \leq x \leq B$ with the variable set-theoretical sum of the N intervals. The hypothesis of (15) is clearly satisfied. The probability that any point x in the interval I shall be contained in the ith interval of length a is clearly $(p/a)a = p$. From this it follows that

$$\Pr\ (x \epsilon X) = p(x) = \begin{cases} 1 - (1 - p)^N \text{ for } 0 \leq x \leq B \\ 0 \text{ elsewhere.} \end{cases} \tag{25}$$

From (15) it follows that

$$E(\mu(X)) = \int_0^B p(x)\, dx = B(1 - (1 - p)^N). \tag{26}$$

(The same formula holds in the case where the N intervals of length a are replaced by N circles of area a and I by a plane domain of area B, provided that for every point of the domain the probability of being contained in the ith circle is equal to a constant p. A similar remark holds for spheres in space.)

To evaluate $E(\mu^2(X))$ in the linear case we make use of the identity

$$\text{(27)}\quad \Pr\,(A \text{ and } B) = \Pr\,(A) + \Pr\,(B) + \Pr\,(\text{neither } A \text{ nor } B) - 1,$$

which holds for any two events A and B. It follows from (27) and (25) that if x and y are any two points of I, then

$$\text{(28)}\quad \begin{aligned} p(x, y) &= \Pr\,(x \in X \text{ and } y \in X) \\ &= \Pr\,(x \in X) + \Pr\,(y \in X) + \Pr\,(x \notin X \text{ and } y \notin X) - 1 \\ &= 1 - 2(1 - p)^N + \Pr\,(x \notin X \text{ and } y \in X). \end{aligned}$$

Let

$$\text{(29)}\quad h(x, y) = \Pr\,(x \notin X \text{ and } y \notin X).$$

Then

$$\text{(30)}\quad h(x, y) = \begin{cases} [1 - (p/a)2a]^N = (1 - 2p)^N, & \text{for } |y - x| \geq a \\ [1 - (p/a)(a + |y - x|)]^N = \left(\dfrac{a - ap - p|y - x|}{a}\right)^N, & \\ & \text{for } |y - x| < a. \end{cases}$$

Now from (21), (28), and (29) we have

$$\text{(31)}\quad \begin{aligned} E(\mu^2(X)) &= \int_0^B \int_0^B p(x, y)\, dy\, dx \\ &= \int_0^B \int_0^B [1 - 2(1 - p)^N + h(x, y)]\, dy\, dx \\ &= B^2[1 - 2(1 - p)^N] + 2 \int_0^B \int_x^B h(x, y)\, dy\, dx. \end{aligned}$$

When the latter integral is evaluated the result is

$$\text{(32)}\quad \begin{aligned} E(\mu^2(X)) = B^2[1 - 2(1 - p)^N] &+ (B - a)^2(1 - 2p)^N \\ &+ \frac{2aB(1 - p)^{N+1}}{(N + 1)p} - \frac{2a(B - a)(1 - 2p)^{N+1}}{(N + 1)p} \\ &- \frac{2a^2}{(N + 1)(N + 2)p^2}[(1 - p)^{N+2} - (1 - 2p)^{N+2}]. \end{aligned}$$

Combining this with (26), we find for the variance of $\mu(X)$ the expression

$$\text{(33)}\quad \begin{aligned} \sigma^2 &= E(\mu^2(X)) - [E(\mu(X))]^2 \\ &= (B - a)^2(1 - 2p)^N - B^2(1 - p)^{2N} + \frac{2aB(1 - p)^{N+1}}{(N + 1)p} \\ &\quad - \frac{2a(B - a)(1 - 2p)^{N+1}}{(N + 1)p} - \frac{2a^2}{(N + 1)(N + 2)p^2}[(1 - p)^{N+2} - (1 - 2p)^{N+2}]. \end{aligned}$$

ON THE DISTRIBUTION OF THE RADIAL STANDARD DEVIATION

By Frank E. Grubbs[1]

Aberdeen Proving Ground

1. Introduction. Of interest in the field of ballistics is a measure of the accuracy of bullets. In acceptance tests of small arms ammunition lots, for example, a sample of rounds from each lot is fired from a fixed rifle at a vertical target placed a specified distance from the rifle. The accuracy of the bullets is taken to be some measure of the scattering (or lack of scattering) of the bullet holes on the target. The purpose of such a test would be to determine whether or not the lot under consideration differs significantly in accuracy from (a) standard values or (b) its predecessors.

One useful measure of accuracy is the radial standard deviation which is defined by the relation

$$Z = \sqrt{\frac{1}{N}\{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2\}}, \tag{1}$$

where x_i and y_i are respectively the abscissa and ordinate of any point measured from an arbitrary origin and N is the sample size.

It will be the purpose of the present discussion to call attention to a series expansion for the distribution of the statistic Z in samples of N assuming that the distribution of all rounds of the lot on the target follow the bivariate normal population law

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{x^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}}, \qquad (x \text{ and } y \text{ statistically independent}) \tag{2}$$

where σ_1^2 and σ_2^2 are the parent variances of x and y respectively. In the above probability density function, the population means are taken to be zero since the statistic Z is quite independent of the origin selected.

2. Moment generating function of Z^2. The distribution of $s_1^2 = \frac{1}{N}\Sigma(x_i - \bar{x})^2$ in samples of N from a normal population is given by the well-known law,

$$dF(s_1^2) = \frac{\frac{N}{2\sigma_1^2}}{\Gamma\left(\frac{N-1}{2}\right)} \left(\frac{Ns_1^2}{2\sigma_1^2}\right)^{\frac{1}{2}(N-3)} e^{-\frac{Ns_1^2}{2\sigma_1^2}} ds_1^2, \qquad s_1^2 \geq 0. \tag{3}$$

The moment generating function of s_1^2 may be found (in a neighborhood of $t = 0$) by straightforward integration:

$$M_{s_1^2}(t) = E(e^{s_1^2 t}) = \int_0^\infty e^{s_1^2 t}\, dF(s_1^2) = \left\{1 - \frac{2\sigma_1^2 t}{N}\right\}^{-\frac{1}{2}(N-1)}. \tag{4}$$

[1] Captain, Ordnance Department, Ballistic Research Laboratory, Aberdeen Proving Ground, Md.

Likewise, for $s_2^2 = \frac{1}{N}\Sigma(y_i - \bar{y})^2$, we have

$$M_{s_2^2}(t) = \left\{1 - \frac{2\sigma_2^2 t}{N}\right\}^{-\frac{1}{2}(N-1)}. \tag{5}$$

Now $M_{Z^2}(t) = M_{s_1^2+s_2^2}(t) = E\{e^{s_1^2t+s_2^2t}\} = E(e^{s_1^2t})\cdot E(e^{s_2^2t})$ since x and y are independent. Thus,

$$M_{Z^2}(t) = M_{s_1^2}(t)\cdot M_{s_2^2}(t) = \left\{1 - \frac{2\sigma_1^2 t}{N}\right\}^{-\frac{1}{2}(N-1)}\left\{1 - \frac{2\sigma_2^2 t}{N}\right\}^{-\frac{1}{2}(N-1)}.$$

3. Distribution function of Z^2. Making use of the Fourier theorem, we have

$$f(Z^2) = \frac{1}{2\pi}\int_{\infty}^{\infty}\left\{1 - \frac{2\sigma_1^2 it}{N}\right\}^{-\frac{1}{2}(N-1)}\left\{1 - \frac{2\sigma_2^2 it}{N}\right\}^{-\frac{1}{2}(N-1)} e^{-iZ^2t}\,dt, \tag{6}$$

at all points of continuity of $f(Z^2)$.

The discussion will be divided preferably into the two cases: Case I: $\sigma_1^2 = \sigma_2^2$, and Case II: $\sigma_1^2 \neq \sigma_2^2$.

Case I: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

In this case the distribution of Z^2 reduces to

$$f(Z^2) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{1 - \frac{2\sigma^2 it}{N}\right\}^{-(N-1)} e^{-iZ^2t}\,dt. \tag{7}$$

It will simplify the algebra to find first the distribution of $u^2 = \frac{NZ^2}{2\sigma^2}$ and then that of Z^2. Since $M_{u^2}(t) = \{1 - t\}^{-(N-1)}$,

$$f(u^2) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\{1 - it\}^{-(N-1)} e^{-iu^2t}\,dt. \tag{8}$$

This integral may be evaluated easily by the calculus of residues since the integrand has only a single pole of order $(N - 1)$ at $t = -i$. We will, however, make use of the following method.

Put $-v = u^2 - iu^2t$; then

$$\begin{aligned} f(u^2) &= \frac{1}{2\pi}\int_{-u^2-i\infty}^{-u^2+i\infty}\left(-\frac{v}{u^2}\right)^{-(N-1)} e^{-v-u^2}\frac{dv}{iu^2} \\ &= \frac{-e^{-u^2}(u^2)^{N-2}}{2\pi i}\int_{-u^2+i\infty}^{-u^2-i\infty} e^{-v}(-v)^{-(N-1)}\,dv. \end{aligned} \tag{9}$$

The integral in the last expression is Hankel's integral [1]; namely,

$$\frac{1}{\Gamma(Z)} = \frac{i}{2\pi}\int_{a+i\infty}^{-a-i\infty} e^{-t}(-t)^{-Z}\,dt, \qquad R(Z) > 0, \qquad a > 0. \tag{10}$$

Therefore $$f(u^2) = \frac{1}{\Gamma(N-1)}e^{-u^2}(u^2)^{N-2},$$

and $$dF(Z^2) = \frac{\dfrac{N}{2\sigma^2}}{\Gamma(N-1)}\, e^{-\frac{NZ^2}{2\sigma^2}} \left(\frac{NZ^2}{2\sigma^2}\right)^{N-2} dZ^2; \text{ from which}$$

$$(11) \qquad dF(Z) = \frac{2\left(\dfrac{N}{2\sigma^2}\right)^{N-1}}{\Gamma(N-1)}\, e^{-\frac{NZ^2}{2\sigma^2}}\, Z^{2N-3}\, dZ.$$

(Note that $f(Z)$ is continuous over $0 \leq Z \leq \infty$.)

This expected result has been obtained by Reno and Mowshowitz [2] who employed an extension of the famous Helmert distribution.

Actually, the result is an obvious one and may be argued as follows: Ns_1^2/σ^2 is distributed as χ^2 with $N - 1$ degrees of freedom and Ns_2^2/σ^2 is also distributed as χ^2 with $N - 1$ degrees of freedom. Hence, the statistic $\frac{N}{\sigma^2}(s_1^2 + s_2^2)$ is, from the additive property of χ^2, distributed like χ^2 with $2N - 2$ degrees of freedom.

We now turn to the general

Case II: $\sigma_1^2 \neq \sigma_2^2$

No generality will be lost by taking $\sigma_1^2 < \sigma_2^2$. In fact, the present attack will hold with obvious modifications provided $\sigma_1^2 < 2\sigma_2^2$.

Recall that

$$(12) \qquad f(Z^2) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left\{1 - \frac{2\sigma_1^2 it}{N}\right\}^{-\frac{1}{2}(N-1)} \left\{1 - \frac{2\sigma_2^2 it}{N}\right\}^{-\frac{1}{2}(N-1)} e^{-iZ^2 t}\, dt,$$

at all continuity points of $f(Z^2)$.

In a manner analogous to that employed by Hsu [3], we replace

$$\left(1 - \frac{2\sigma_2^2 it}{N}\right) \text{ by } \frac{\sigma_2^2}{\sigma_1^2}\left(1 - \frac{2\sigma_1^2 it}{N}\right)\left\{1 - \frac{1 - \sigma_1^2/\sigma_2^2}{1 - 2\sigma_1^2 it/N}\right\}.$$

Further, since

$$\left|\frac{1 - \sigma_1^2/\sigma_2^2}{1 - 2\sigma_1^2 it/N}\right| < 1,$$

we may write

$$\left\{1 - \frac{2\sigma_2^2 it}{N}\right\}^{-\frac{1}{2}(N-1)} = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}(N-1)} \left(1 - \frac{2\sigma_1^2 it}{N}\right)^{-\frac{1}{2}(N-1)} \sum_{r=0}^{\infty} \frac{\Gamma\left(\dfrac{N-1}{2} + r\right)}{\Gamma\left(\dfrac{N-1}{2}\right)\Gamma(r+1)} \cdot \frac{\left(1 - \dfrac{\sigma_1^2}{\sigma_2^2}\right)^r}{\left(1 - \dfrac{2\sigma_1^2 it}{N}\right)^r}.$$

Thus,

$$(13)\qquad f(Z^2) = \frac{\left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}(N-1)}}{2\pi} \int_{-\infty}^{\infty} \sum_{r=0}^{\infty} \frac{\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^r}{r\beta\left(\frac{N-1}{2}, r\right)} \left\{1 - \frac{2\sigma_1^2 it}{N}\right\}^{-(N+r-1)} e^{-iZ^2t}\, dt,$$

with the understanding that $r\beta\left(\frac{N-1}{2}, r\right) = 1$ for $r = 0$.

We note that the moduli of the terms of the above series are for all t not greater than the corresponding terms of the following convergent series of positive terms:

$$\sum_{r=0}^{\infty} \frac{\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^r}{r\beta\left(\frac{N-1}{2}, r\right)}.$$

Therefore, uniform convergence over $(-\infty, \infty)$ is established. To show that we may integrate over the infinite interval term by term, we observe that $| S(t) - S_r(t) | \leq \epsilon\varphi(t)$ for all t and all large r, where

$$S(t) = \left\{1 - \frac{2\sigma_1^2 it}{N}\right\}^{-\frac{1}{2}(N-1)} \left\{1 - \frac{2\sigma_2^2 it}{N}\right\}^{-\frac{1}{2}(N-1)},$$

$S_r(t) =$ the sum of the first $r + 1$ terms of the series, and the function $\varphi(t) = \left| 1 - \frac{2\sigma_1^2 it}{N}\right|^{-2}$ which is integrable over $(-\infty, \infty)$. That is, $S_r(t)$ converges to $S(t)$ uniformly relative to $\varphi(t)$.[2] Hence,

$$(14)\qquad f(Z^2) = \frac{\left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}(N-1)}}{2\pi} \sum_{r=0}^{\infty} \frac{\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^r}{r\beta\left(\frac{N-1}{2}, r\right)} \int_{-\infty}^{\infty} \left\{1 - \frac{2\sigma_1^2 it}{N}\right\}^{-(N+r-1)} e^{-iZ^2t}\, dt.$$

We have already carried out the integration under Case I with the exception that $(N - 1)$ should now be replaced by $(N + r - 1)$. The distribution of Z^2 will then be given by

$$(15)\qquad dF(Z^2) = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}(N-1)} \sum_{r=0}^{\infty} \frac{\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^r}{r\beta\left(\frac{N-1}{2}, r\right)} \frac{\frac{N}{2\sigma_1^2}}{\Gamma(N + r - 1)} \cdot e^{-\frac{NZ^2}{2\sigma_1^2}} \left(\frac{NZ^2}{2\sigma_1^2}\right)^{N+r-2} d(Z^2).$$

[2] The author is indebted to Prof. E. J. McShane for this definition which is due to Prof. E. H. Moore. It may be shown easily that $\lim_{r\to\infty} \int_{-\infty}^{\infty} S_r(t)\, dt = \int_{-\infty}^{\infty} \lim_{r\to\infty} S_r(t)\, dt$.

Finally, the distribution function of Z is

$$(16)\quad dF(Z) = 2\left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}(N-1)} \sum_{r=0}^{\infty} \frac{\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^r}{r\beta\left(\frac{N-1}{2}, r\right)} \cdot \frac{\left(\frac{N}{2\sigma_1^2}\right)^{N+r-1}}{\Gamma(N+r-1)} e^{-\frac{NZ^2}{2\sigma_1^2}} Z^{2N+2r-3}\, dz.$$

We remark that the above series expansion holds, of course, for N odd or even. In case N is odd it may be shown that the distribution function may be expressed as a finite series of Incomplete Gamma Functions.[3] However, the finite expansion for N odd appears to offer no marked advantage since for computational purposes the infinite series expansion converges quite rapidly (N either odd or even) and may be put into a convenient form given below.

4. Computational form for the distribution function. In deciding whether or not an observed value of Z is significant and likewise in control chart procedure, one is interested in the percentage points of $f(Z)$. For example, it may be desired to find the value of k such that $P\{Z \leq k\sqrt{\sigma_1^2 + \sigma_2^2}\} = .995$, say, for various sample sizes N. In this connection it will be convenient to work with the distribution of Z^2, for $P\{Z \leq k\sqrt{\sigma_1^2 + \sigma_2^2}\} = P\{Z^2 \leq k^2(\sigma_1^2 + \sigma_2^2)\}$ also. Now,

$$P\{Z^2 \leq k^2(\sigma_1^2 + \sigma_2^2)\} = \int_0^{k^2(\sigma_1^2+\sigma_2^2)} dF(Z^2)$$

$$(18)\qquad = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}(N-1)} \sum_{r=0}^{\infty} \frac{\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^r}{r\beta\left(\frac{N-1}{2}, r\right)} \cdot \frac{\left(\frac{N}{2\sigma_1^2}\right)}{\Gamma(N+r-1)} \cdot \int_0^{k^2(\sigma_1^2+\sigma_2^2)} e^{-\frac{NZ^2}{2\sigma_1^2}} \left(\frac{NZ^2}{2\sigma_1^2}\right)^{N+r-2} d(Z^2),$$

since we may integrate the series term by term over the entire range of Z^2 or any part of it [5]. In the terminology of Karl Pearson's Incomplete Gamma Function [3],

$$(19)\qquad I(u, p) = \frac{1}{\Gamma(p+1)} \int_0^{u\sqrt{p+1}} e^{-v} v^p\, dv,$$

we may write the above series in the form

$$P\{Z^2 \leq k^2(\sigma_1^2 + \sigma_2^2)\}$$

$$(20)\qquad = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\frac{1}{2}(N-1)} \sum_{r=0}^{\infty} \frac{\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^r}{r\beta\left(\frac{N-1}{2}, r\right)} I\left\{\frac{Nk^2\left(1 + \frac{\sigma_2^2}{\sigma_1^2}\right)}{2\sqrt{N+r-1}}, N+r-2\right\}.$$

[3] Prof. C. C. Craig kindly pointed out this fact to the author.

It is indeed convenient and enlightening that the result is a function of the ratio, σ_1^2/σ_2^2, and not σ_1^2 and/or σ_2^2 explicitly.

Hence, for a given sample size and ratio of σ_1^2/σ_2^2, we may find k by inverse interpolation such that $P\{Z \leq k\sqrt{\sigma_1^2 + \sigma_2^2}\} = \alpha$, any desired level of probability.

5. Moments and percentage points for Case I. For the case met many times in practice, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we will give a table of the mean and standard deviation and also several probability levels which are obtainable directly from the percentage points of the χ^2 distribution [6].

From (11), we have

$$E(Z^k) = \frac{2\left(\frac{N}{2\sigma^2}\right)^{N-1}}{\Gamma(N-1)} \int_0^\infty e^{-\frac{NZ^2}{2\sigma^2}} Z^{2N+k-3}\, dZ \tag{21}$$

$$= \frac{\Gamma(N-1+k/2)}{\Gamma(N-1)} \left(\frac{2\sigma^2}{N}\right)^{\frac{1}{2}k}.$$

Thus,

$$\mu'_{1:z} = \frac{\Gamma(N-1/2)}{\Gamma(N-1)} \sqrt{\frac{2}{N}}\,\sigma, \tag{22}$$

$$\mu'_{2:z} = \frac{2(N-1)}{N}\,\sigma^2, \tag{23}$$

and

$$\mu_{2:z} = \frac{2}{N}\left\{N - 1 - \left[\frac{\Gamma(N-1/2)}{\Gamma(N-1)}\right]^2\right\}\sigma^2. \tag{24}$$

In the table below, the mean and standard deviation are given as a multiple of $\sqrt{2}\sigma$, and $k_{.95}$, for example, is that value of k such that $P\{Z \leq k\sqrt{2}\sigma\} = .95$.

TABLE I

N	Mean	Standard Deviation	Percentage Points			
			$k_{.005}$	$k_{.05}$	$k_{.95}$	$k_{.995}$
2	.6267	.3276	.0501	.1602	1.2239	1.6276
3	.7675	.2786	.1857	.3442	1.2575	1.5738
4	.8308	.2443	.2906	.4521	1.2546	1.5226
5	.8670	.2198	.3667	.5227	1.2453	1.4817
6	.8904	.2014	.4239	.5730	1.2351	1.4488
7	.9068	.1869	.4686	.6110	1.2255	1.4218
8	.9189	.1752	.5046	.6408	1.2167	1.3991
9	.9282	.1653	.5345	.6651	1.2087	1.3798
10	.9356	.1569	.5597	.6852	1.2014	1.3630
11	.9416	.1498	.5813	.7023	1.1949	1.3483
12	.9466	.1434	.6001	.7170	1.1889	1.3353
13	.9508	.1378	.6166	.7298	1.1835	1.3237
14	.9544	.1330	.6313	.7411	1.1784	1.3132
15	.9575	.1285	.6445	.7512	1.1738	1.3038

REFERENCES

[1] Whittaker and Watson, *Modern Analysis*, Fourth Edition, 1922, pp. 244–246.

[2] F. V. Reno and Simon Mowshowitz, "The distribution of the radial standard deviation", Ballistic Research Laboratory Report No. 322, Aberdeen Proving Ground, Md.

[3] P. L. Hsu, "Contribution to the theory of student's *t*-Test as applied to the problem of two samples", *Statistical Research Memoirs*, Vol. 2, December 1938, pp. 1–24.

[4] Karl Pearsôn (Editor), *Tables of the Incomplete Gamma Function*, Cambridge University Press, 1934.

[5] Bromwich, *An Introduction to the Theory of Infinite Series*, MacMillan, 1908, pp. 453–454.

[6] Catherine M. Thompson, "Table of percentage points of the χ^2 distribution", *Biometrika*, Vol. 32, Part II (1941), pp. 187–191.

A MATRIX PRESENTATION OF LEAST SQUARES AND CORRELATION THEORY WITH MATRIX JUSTIFICATION OF IMPROVED METHODS OF SOLUTION

By Paul S. Dwyer

University of Michigan

1. Introduction and summary. It is the aim of this paper to exhibit, by using elementary matrix theory, the basic concepts of least squares and correlation theory, the solution of the normal equations, and the presentation and justification of recently developed and newly proposed techniques into a single, compact, and short presentation. We shall be mainly concerned with the following topics:

a. Basic least squares theory including derivation of normal equations, the theoretical solution of these equations (regression coefficients), the standard errors of these solutions, and the standard error of estimate.

b. The more specific theory (correlation theory) resulting from applying the general least squares results to the standardized distributions.

c. A matrix presentation of the Doolittle solution.

d. A simple matrix justification of methods, previously presented, for getting least squares and multiple correlation constants from the entries of an abbreviated Doolittle solution.

e. A presentation of a more general theory which the matrix presentation reveals.

f. The outline of a "square root" method as an alternative to theDoolittle method.

The reader should be familiar with elementary matrix theory such as that outlined on pages 1–57 of Aitken's book [1].

No previous knowledge of the Doolittle technique is demanded although a familiarity with the notation and contents of two earlier papers [2], [3] is advised, particularly for those who are interested in the computational aspects.

The presentation here is theoretical and is not concerned with such computational topics as the number of decimal places required, etc. With reference to the number of places, the reader is referred to the recent paper of Professor Hotelling [4].

2. Notation. Let $[x'_{ij}]$ with $1 \leq i \leq N$ and $1 \leq j \leq n$ be the n by N matrix of observed variates of n "predicting variables" for N individuals with i indicating the individual and j the variable. Let $[y'_i]$ be the one by N column matrix of the observed variates of the "predicted" variable. Let the matrices of deviations from the variable means be indicated by $[x_{ij}] = X$ and $[y_i] = Y$. Then by the least squares hypothesis we are to find numbers $b_{y1\cdots\cdot}$, $b_{y2\cdots\cdot}$, $\cdots$, $b_{yn\cdots\cdot}$ such that

$$e_i = y_i - (x_{i1}b_{y1\cdots\cdot} + x_{i2}b_{y2\cdots\cdot} + \cdots + x_{in}b_{yn\cdots\cdot}),$$

shall have a minimum variation (standard deviation). We then denote the $b_{yi\cdots}$ by the one by n column matrix B and the e_i by the one by N column matrix E and have

$$E = Y - XB, \tag{1}$$

as the basic matrix equation.

It may be noted further that the fitted values of y_i are given by the one by N matrix product $XB = \underline{Y}$. Using this notation (1) appears as

$$E = Y - \underline{Y}. \tag{1$'$}$$

3. Basic least squares theory.

a. Sum of squares of residuals. The condition for minimum variation in this situation (variates measured from means) is equivalent to the condition for minimum sum of squares of residuals. In matrix notation this sum of the squares of the residuals can be written

$$E'E \text{ with } E = Y - XB = Y - \underline{Y}, \quad \text{and } E' \text{ the transpose of } E. \tag{2}$$

b. The normal equations. Differentiating (2) with respect to B' we find the necessary condition to be

$$X'E = 0. \tag{3}$$

This matrix equation gives the normal equations in implicit form. More explicitly by (1) we have $X'(Y - XB) = 0$ so

$$X'XB = X'Y. \tag{4}$$

The reader should immediately recognize that (4) is the matrix equivalent of the usual statement of the normal equations where deviations from the means are used. It should be noted also that (3) and (4) can be written in the form

$$X'\underline{Y} = X'Y \quad \text{from whence at once } \underline{Y}'\underline{Y} = \underline{Y}'Y. \tag{5}$$

c. Solution of normal equations. The theoretical solution of (4) is accomplished at once and results in

$$B = (X'X)^{-1}X'Y = (X'X)^{-1}X'\underline{Y}. \tag{6}$$

d. Standard deviation of residuals. The standard deviation of residuals is $\sqrt{E'E/N}$.

In order to evaluate this we note that

$$\underline{Y}'E = B'X'E = 0, \qquad \text{and } \underline{Y}'Y = \underline{Y}'\underline{Y}, \tag{7}$$

Thus

$$E'E = (Y - \underline{Y})'E = Y'E = Y'(Y - XB) = Y'Y - Y'XB, \tag{8}$$

and

$$Y'XB = Y'\underline{Y} = \underline{Y}'\underline{Y} = \underline{Y}'Y. \tag{9}$$

Since $Y'Y = \Sigma y^2$, we have

$$E'E = \Sigma y^2 \left(1 - \frac{Y'XB}{\Sigma y^2}\right), \tag{10}$$

so that, dividing by N and taking the square root

$$s_e = s_y \sqrt{1 - \frac{Y'XB}{\Sigma y^2}}. \tag{11}$$

If the relation between the estimated standard deviations in the population is desired then divide each side of (10) by the number of degrees of freedom and get

$$\sigma_e = \sigma_y \sqrt{1 - \frac{Y'XB}{\Sigma y^2}}. \tag{12}$$

Alternative formulas to (11) and (12) are obtained by replacing $Y'XB$ by its equivalent expressions in (9).

e. Formulas for multiple correlation coefficient. It is to be noted that the numerical quantity $Y'XB/\Sigma y^2$ plays an important role in measuring the ratio σ_e/σ_y. It is customary to use this quantity as the definition of the square of the multiple correlation coefficient so we have

$$r^2_{y \cdot x_1 \cdots x_n} = \frac{Y'XB}{\Sigma y^2} = \frac{B'X'XB}{\Sigma y^2} = \frac{Y'X(X'X)^{-1}X'Y}{\Sigma y^2} = \frac{\underline{Y}'Y}{\Sigma y^2} = \frac{\sigma^2_{\underline{y}}}{\sigma^2_y}. \tag{13}$$

f. Formulas for correlation coefficient. When $n = 1$, $X'X = \Sigma x^2$, $Y'X = X'Y = \Sigma XY$, $B = b$ and (13) gives

$$r_{yx} = \sqrt{\frac{b\Sigma xy}{\Sigma y^2}} = b\sqrt{\frac{\Sigma x^2}{\Sigma y^2}} \left(= b\,\frac{\sigma_x}{\sigma_y}\right) = \frac{\Sigma xy}{\sqrt{\Sigma x^2\, \Sigma y^2}} = \frac{\sigma_{\underline{y}}}{\sigma_y}. \tag{14}$$

Many of the above developments can be duplicated, without formal use of matrix theory, by judicious use of symbolism and substitution. See for example the presentations of Kirkham [5], Bacon [6], and Guttman [7].

g. Errors of regression coefficients. If B_0 is an approximation to B such that $B_0 + \Delta B = B$ then (6) can be written

$$B_0 + \Delta B = (X'X)^{-1}X'Y$$

and

$$\Delta B = (X'X)^{-1}X'(Y - XB_0). \tag{15}$$

This formula can be used in finding corrections ΔB necessary to change any proposed trial solution, B_0, into a correct solution. It could also be used in extending the accuracy of a solution after an approximation had been secured to a specific number of places. It has greater utility however in another problem.

We suppose that the predicting variables, the x's, contain no errors but that there are errors in the observed values of y. Let the hypothetical observed values of y be indicated by Y and the recorded observed values of y by Y_0. Let the values of B_0 be the regression coefficients obtained by using the recorded observed values Y_0. Thus $\Delta B = 0$ when Y is replaced by Y_0 in (15). Now let $Y - XB_0$, the "true" residual errors of the recorded observed values, be indicated by E. Then (15) becomes

$$\Delta B = (X'X)^{-1}X'E. \tag{16}$$

Sampling theory can be applied to (16) to obtain a formula for the standard error of the regression coefficient. It is assumed that the "true" residual errors are independent with a common standard deviation σ_e. The values of ΔB are then linear functions of these errors. It follows that

$$\sigma_B^2 = \sigma_{\Delta B}^2 = (X'X)^{-1}X'X(X'X)^{-1}\sigma_e^2 = (X'X)^{-1}\sigma_e^2. \tag{17}$$

The standard errors of the regression coefficients are thus formed by multiplying σ_e by the square roots of the diagonal terms of the inverse of $X'X$.

4. Standard variates. Use of correlation matrix. Many of the formulas of section 3 are simplified with the use of some type of standardization. In particular it is possible to reduce the matrix $X'X$ to the matrix R of correlation coefficients by replacing x by t_x/N where $t_x = x/s$. If y is similarly replaced and B by $\mathbf{B}$, then $X'Y = R_{xy}$ and $Y'X = R'_{xy}$, $Y'Y = \Sigma y^2 = 1$ and selected formulas from section 3 become

$$R\mathbf{B} = R_{xy} \tag{18}$$

$$\mathbf{B} = R^{-1}R_{xy} \tag{19}$$

$$r^2_{y.x_1\dots x_n} = R'_{xy}\mathbf{B} = \mathbf{B}'\mathbf{R}\mathbf{B} = R'_{xy}R^{-1}R_{xy}. \tag{20}$$

Classical multiple correlation formulas, determinantal and otherwise, are "covered" by the matrix formulas (20).

5. Matrix presentation of a Doolittle solution. Least squares and correlation constants can also be obtained from the entries of a Doolittle solution. We first outline a matrix description of the Doolittle solution of the equation $AX = G$ with $A = [a_{ij}]$ symmetric and of order n.

Let S_1 be a (n by n) matrix with the first row composed of the elements a_{1j} and all other elements 0. Let T_1 be a similar matrix with first row elements $b_{1j} = a_{1j}/a_{11}$ and all other elements 0. Then $A - S_1'T_1 = A_1 = [a_{ij.1}]$ is a symmetric (n by n) matrix with all elements of the first row and the first column 0.

Next let S_2 be a (n by n) matrix with second row elements $a_{2j.1} = a_{2j} - a_{21}b_{1j}$ and all other elements 0. Let T_2 be a (n by n) matrix with second row elements $b_{2j.1} = a_{2j.1}/a_{22.1}$ and all other elements 0. Then it follows that the matrix

$A_1 - S_2'T_2 = [a_{ij.12}]$ is a symmetric (n by n) matrix with the elements of the first 2 columns and the first 2 rows all 0.

This process is continued through successive steps, an additional row and column being made identically 0 at each step, through n steps. At the end of n steps we have the result.

$$A - S_1'T_1 - S_2'T_2 - \cdots S_n'T_n = 0. \tag{21}$$

This development, when applied to each side of the matrix equation, provides the basis for an equation solving technique which Aitken has called the "method of pivotal condensation" (8) but which the author feels is more adequately characterized as the "method of single division" (9). The Abbreviated Doolittle method can be obtained as an abbreviation of this method. It is not necessary to compute all the elements of the successive matrices $A_1A_2 \cdots$, etc. but only the non-zero elements of the S_1, T_1, S_2, $T_2 \cdots$ etc. matrices.

Consider the so called triangular matrix $S = S_1 + S_2 + S_3 + \cdots + S_n$ with its rows composed of the non-zero rows of the S_j. Consider also the matrix $T = T_1 + T_2 + \cdots + T_n$. Then

$$S'T = S_1'T_1 + S_2'T_2 + \cdots + S_n'T_n \tag{22}$$

$$\text{since } S_i'T_j = 0 \text{ when } i \neq j.$$

It follows that (21) can be written

$$A - S'T = 0. \tag{23}$$

An efficient way of building up these matrices S and T in practice and in making the corresponding transformations on the right side of the equation is the Abbreviated Doolittle method. It is apparent from (23) that the Doolittle method is directed, in part at least, toward the factorization of the symmetric matrix A into two triangular matrices.

It should be noted that these triangular matrices are related by the matrix formula

$$S = DT, \tag{24}$$

where D is the diagonal matrix with diagonal elements a_{11}, $a_{22 \cdot 1}$, $a_{33 \cdot 12}$, $\cdots$, $a_{nn \cdot 123 \cdots n-1}$.

Operations performed on the left of the matrix equations $AX = G$ are also performed on the right side so that the Doolittle technique results in the establishment of the auxiliary matrix equations.

$$SX = SA^{-1}G. \tag{25}$$

$$TX = TA^{-1}G. \tag{26}$$

A simple outline ($n = 3$) of the form of the Abbreviated Doolittle method is presented for the purpose of identifying these matrices. A is symmetric and G is the column matrix $[a_{i4}]$.

$$\begin{array}{ccc|c}
a_{11} & a_{12} & a_{13} & a_{14} \\
- & a_{22} & a_{23} & a_{24} \\
- & - & a_{33} & a_{34} \\
\hline
a_{11} & a_{12} & a_{13} & a_{14} \\
1 & b_{12} & b_{13} & b_{14} \\
\hline
 & a_{22\cdot 1} & a_{23\cdot 1} & a_{24\cdot 1} \\
 & 1 & b_{23\cdot 1} & b_{24\cdot 1} \\
\hline
 & & a_{33\cdot 12} & a_{34\cdot 12} \\
 & & 1 & b_{34\cdot 12}
\end{array}$$

The matrix S is then $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22\cdot 1} & a_{23\cdot 1} \\ 0 & 0 & a_{33\cdot 12} \end{bmatrix}$, the matrix T is

$$\begin{bmatrix} 1 & b_{12} & b_{13} \\ 0 & 1 & b_{23\cdot 1} \\ 0 & 0 & 1 \end{bmatrix}, \qquad SA^{-1}G \text{ is } \begin{bmatrix} a_{14} \\ a_{24\cdot 1} \\ a_{34\cdot 12} \end{bmatrix}, \qquad TA^{-1}G \text{ is } \begin{bmatrix} b_{14} \\ b_{24\cdot 1} \\ b_{34\cdot 12} \end{bmatrix}.$$

6. Least squares and multiple correlation constants from the Doolittle solution. The inverse of A is needed for many formulas. We set up a technique for solving $AY = \mathrm{I}$ simultaneously with $AX = G$. This is indicated symbolically by

$$\begin{array}{c|c|c}
A & G & I \\
\hline
S & SA^{-1}G & SA^{-1} \\
\hline
T & TA^{-1}G & TA^{-1}
\end{array}.$$

It follows at once that

$$(SA^{-1})'(TA^{-1}) = A^{-1}S'TA^{-1} = A^{-1}AA^{-1} = A^{-1}. \tag{27}$$

This matrix multiplication is easily and readily accomplished when the matrices are in the Doolittle form.
Similarly

$$(SA^{-1})'(TA^{-1}G) = A^{-1}G = X, \text{ the matrix of solutions of } AX = G \text{ and} \tag{28}$$

$$(SA^{-1}G)'(TA^{-1}G) = G'A^{-1}G. \tag{29}$$

It is interesting to note further that $(SA^{-1})'$ and T are inverse triangular matrices since

$$(SA^{-1})'T = A^{-1}S'T = I. \tag{30}$$

S' and TA^{-1} have a similar relationship.

In the case of least squares theory $A = X'X$, $G = X'Y$, $X = B$ so that the formulas (27)(28)(29) become

(31) $(SA^{-1})'(TA^{-1}) = A^{-1} = (X'X)^{-1}$.

(32) $(SA^{-1})'(TA^{-1}G) = X = B$.

(33) $$(SA^{-1}G)'(TA^{-1}G) = G'A^{-1}G = Y'X(X'X)^{-1}X'Y$$
$$= Y'XB = B'X'XB = Y'Y.$$

If the normal equations are reduced to standard form $A = R$, $X = \mathbf{B}$, $G = R_{xy}$ and we have

(34) $(SA^{-1})'(TA^{-1}) = A^{-1} = R^{-1}$.

(35) $(SA^{-1})'(TA^{-1}G) = X = \mathbf{B}$.

(36) $(SA^{-1}G)'(TA^{-1}G) = G'A^{-1}G = R'_{xy}R^{-1}R_{xy} = R'_{xy}\mathbf{B} = \mathbf{B}'R\mathbf{B} = r^2_{y\cdot x_1\cdots x_n}$.

The reader is referred to an earlier paper [3, 457] for an illustration of these techniques.

It should be noted that the solution is a cumulative one in the sense that solutions involving n predicting variables are obtained from solutions involving $n - 1$ predicting variables by the addition of paired products. This is a highly desirable feature as it makes possible direct analyses showing the effect of an added predicting variable.

7. A more general theory—solution of matrix equations by factorization. Examination of the results of section 5 leads one at once to a consideration of a more general theory. The key formula in this development is $A - S'T = 0$ and all subsequent formulas stem from this. Hence if A can be factored into any matrices, S' and T, not necessarily triangular, the results of section 6 follow.

From a practical standpoint it is desired that the factorization process yield, simultaneously, the values S, T, $SA^{-1}G$; $TA^{-1}G$; SA^{-1} and TA^{-1} as the Doolittle method does. But, formally, these can be computed if S and T are known.

8. A "square root" method. A most interesting and practical special case of the above method is that in which the triangular matrices S and T are equal. It appears that a technique based on this property would have some advantages over the Doolittle method since the double rows of the Doolittle solution could be replaced by single rows, while the formulas of sections 5 and 6 are just as applicable. Now such a technique is easily devised. From (23) and (24) we see that

(37) $$A - S'D^{-1}S = 0,$$

where D is a diagonal matrix.
We replace $D^{-\frac{1}{2}}S$ by a new S, $(\mathrm{D}^{-\frac{1}{2}}S)'$ by a new S' and have

(38) $$A - S'S = 0.$$

The technique of solution is similar to that of the Doolittle except that the entries $s_{ij\cdots}$ are $a_{ij\cdots}/\sqrt{a_{jj\cdots}}$. These values $s_{ij\cdots}$ are thus geometric means of the values $a_{ij\cdots}$ and $b_{ij\cdots}$.

A simple machine technique is available for computing these entries. In some respects the solution is superior to the Doolittle solution. It is hardly pertinent to the subject matter of this paper to present a detailed discussion of the merits of this method, with the numerical illustrations. This will be done in a later paper.

After arriving at this method by the steps described above, it seemed surprising that such a simple and compact method has not been discovered by some previous worker. Although matrix factorization is not a new subject, I have not found evidence that it has been utilized so directly in the problem of solving matrix equations. The nearest approach I have discovered is the paper by Banachiewicz [10], in which a "square root" method is used in factoring A.

REFERENCES

[1] A. C. Aitken, *Determinants and Matrices*, Oliver and Boyd, London, 1942.

[2] P. S. Dwyer, "The Doolittle technique," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 449–458.

[3] P. S. Dwyer, "Recent developments in correlation technique," *Jour. Amer. Stat. Assn.* Vol., 37 (1942), pp. 441–460.

[4] Harold Hotelling, "Some new methods in matrix calculation," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 1–34.

[5] W. Kirkham, "Note on the derivation of the multiple correlation coefficient," *Annals of Math. Stat.*, Vol. 8 (1937), pp. 68–71.

[6] H. M. Bacon, "Note on a formula for the multiple correlation coefficient," *Annals of Math. Stat.* Vol. 9, (1938), pp. 227–229.

[7] Louis Guttman, "A note on the derivation of a formula for multiple and partial correlation," *Annals of Math. Stat.*, Vol. 9 (1938), pp. 305–308.

[8] A. C. Aitken "Studies in practical mathematics I. The evaluation, with application, of a certain triple product matrix," *Proc. Royal Soc., Edinburgh*, Vol. 57 (1937), pp. 172–181.

[9] P. S. Dwyer, "The solution of simultaneous equations," *Psychometrika*, Vol. 6 (1941), pp. 101–129.

[10] T. Banachiewicz, "An outline of the Cracovian algorithm of the method of least squares," *Astr. Jour.* Vol., 50 (1942), pp. 38–41.

ON THE STATISTICS OF SENSITIVITY DATA

By Benjamin Epstein and C. West Churchman

Frankford Arsenal

1. Introduction. "Sensitivity data" is a general term for that type of experimental data for which the measurement at any point in the scale destroys the sample; as a consequence, new samples are required for each determination. Examples of such data occur in biology in dosage-mortality determinations, in psychophysics in questions concerning sensitivity responses, and, more recently, in the theory of solid explosives, in questions concerning the sensitivity of explosive or detonative mixtures.

Methods of analyzing such data have been discussed by Bliss[1] and Spearman[2], and others. The present paper is a generalization of Spearman's result; it is the feeling of the authors that Spearman's method, if properly founded in mathematical theory, is preferable to Bliss', for it does not necessitate the assumption of some type of distribution prior to analysis, and hence resembles the standard treatment of independent observations made on the same object.

Throughout the following discussion, we let x_i be the magnitude of a certain "stimulus" (be it dosage, physical stimulus, or strength of blow) and p_i the corresponding fraction of objects unaffected by the stimulus. Bliss' method consisted in assuming that the p_i represented the cumulative distribution of some known function (in his case, the normal function), and hence the p_i could be transformed into a variable t_i linearly dependent on the x_i. The difficulty of this treatment, in addition to the distribution assumption, lies in the fact that the t_i do not have equal standard errors, and the straight line fit is very cumbersome.

Instead, Spearman makes the much simpler assumption that if p_i is unaffected at x_i, and p_{i+1} at x_{i+1}, then $p_i - p_{i+1}$ is an estimate of the fraction that is just affected (i.e., the fraction of those that have "critical" responses) at about $\frac{1}{2}(x_i + x_{i+1})$. If the x_i are evenly spaced, as we shall assume them to be throughout, and $p_1 = 1.0$ and $p_n = 0$, then any set of sensitivity data may be transformed into a set of data on critical responses classified into classes whose midpoints are evenly spaced. Without loss of generality, we shall assume the x_i's to be integers and the intervals to be unity. The data on critical responses can then be treated in the normal.way, and $\bar{X}$ and all the measures of dispersion calculated in the usual fashion. In order to justify such procedures, however, it is necessary to show how the sampling errors of $\bar{X}$ and the higher moments can be estimated.

[1] C. I. Bliss, "The calculation of the dosage mortality curve," *Annals of Applied Biology*, Vol. 22, pp. 134–167.

[2] C. Spearman, "The method of 'right and wrong cases' (constant stimuli) without Gauss' formulae," *British Jour. of Psych.*, Vol. 2, 1908, pp. 227–242.

2. The moments and their errors. By definition,

$$\bar{X} = \sum_{i=1}^{n} (p_i - p_{i+1})(x_i + x_{i+1})/2 = \sum_{i=1}^{n} (p_i - p_{i+1})(x_i + \tfrac{1}{2}). \tag{1}$$

If we let x_1 represent the stimulus for which none of the samples can be affected, then

$$\bar{X} = x_1 + .5 + \sum_{i=2}^{n-1} p_i , \tag{2}$$

as Spearman has shown (3). Since x_1 is constant, and the p_i are all independent (non-correlated), it follows that (N_i being the number of objects in the ith sample)

$$\sigma^2_{\bar{X}} = \sigma^2_{\Sigma p_i} = \sigma^2_{p_1} + \sigma^2_{p_2} + \cdots + \sigma^2_{p_n} = \sum \sigma^2_{p_i} = \sum_{i=2}^{n-1} \frac{p_i q_i}{N_i} \tag{3}$$

(since $\sigma^2_{p_1} = \sigma^2_{p_n} = 0$).

Again by definition, the qth moment about the origin is

$$\mu'_q = \sum_{i=1}^{n} (p_i - p_{i+1})(x_i + \tfrac{1}{2})^q. \tag{4}$$

As before $x_1 + .5$ can be taken as the origin ($x_1 + .5 = 0$), in which case we have

$$\begin{aligned} \mu'_q = (p_1 - p_2)\cdot 0^q + (p_2 - p_3)\cdot 1^q + (p_3 - p_4)\cdot 2^q \\ + \cdots + (p_{n-1} - p_n)(n - 1)^q. \end{aligned} \tag{5}$$

If we let $b_{q,i}$ represent the ith first difference of the consecutive qth powers of the positive integers (including 0), then

$$\mu'_q = \sum_{i=2}^{n-1} b_{q,i} p_i , \tag{6}$$

by expansion of (5). Hereafter all Σ will be taken from $i = 2$ to $i = n - 1$.

Evidently

$$\sigma^2_{\mu'_q} = \sum_{i=2}^{n-1} b^2_{q,i} \left(\frac{p_i q_i}{N_i}\right), \tag{7}$$

or

$$\sigma^2_{\mu'_q} = \sum_{i=2}^{n-1} b^2_{q,i} \sigma^2_{p_i} . \tag{8}$$

We are interested now in the standard error of the qth moment about the sample mean. To obtain this, compute first the correlation between the qth and rth moments about the origin.

If $\delta\mu_q'$ is taken to be the sample error in μ_q' due to deviations δp_i from the true values, then we have

$$\delta\mu_q' = \sum_{i=2}^{n-1} b_{q,i}\,\delta p_i$$

$$\delta\mu_r' = \sum_{i=2}^{n-1} b_{r,i}\,\delta p_i .$$

Hence

$$\delta\mu_q'\,\delta\mu_r' = \sum b_{q,i}\,b_{r,i}(\delta p_i)^2 + \sum_{i\neq j}(b_{q,i}\,b_{r,j} + b_{q,j}\,b_{r,i})\,\delta p_i\,\delta p_j .$$

Summing for all samples:

$$\begin{aligned}\sigma_{\mu_q'}\sigma_{\mu_r'}r_{\mu_q'\mu_r'} &= \sum b_{q,i}\,b_{r,i}\,\sigma^2_{p_i} + \sum_{i\neq j}(b_{q,i}\,b_{r,j} + b_{q,j}\,b_{r,i})(\sigma_{p_i}\sigma_{p_j}r_{p_ip_j}) \\ &= \sum b_{q,i}\,b_{r,i}\,\sigma^2_{p_i} .\end{aligned} \tag{9}$$

Since evidently $r_{p_ip_j}$ vanishes for all $i \neq j$ (the p_i being completely independent in the statistical sense).

In particular, when $\mu_r' = \mu_1' = \bar{X}$, we have

$$\sigma_{\mu_q'}\sigma_{\bar{X}}r_{\mu_q',\bar{X}} = \Sigma b_{1,i}\sigma^2_{p_i} . \tag{10}$$

By definition, the qth moment about the mean will be

$$\mu_q = \Sigma(p_i - p_{i+1})(x_i + \tfrac{1}{2} - \bar{X})^q = \Sigma p_i'(x_i + \tfrac{1}{2} - \bar{X})^q \tag{11}$$

where $p_i' = p_i - p_{i+1}$.

For computational purposes, this may be written as

$$\mu_q = \mu_q' - q\bar{X}\mu_{q-1}' + \frac{q(q-1)}{2}\bar{X}^2\mu_{q-2}' + \cdots + {}_qC_r\bar{X}^r\mu_{q-r+1} + \cdots + \bar{X}^q \tag{12}$$

where $\bar{X} = \Sigma p_i = \mu_1'$, if $x_1 + \frac{1}{2}$ is the origin.

To obtain $\sigma^2_{\mu_q}$, where $\bar{X}$ is estimated from the sample, we may follow the usual procedures, arguing that

$$\delta\mu_q = \Sigma\{(x_i + \tfrac{1}{2})^q\delta p_i'\} - q\delta\bar{X}\Sigma(x_i + \tfrac{1}{2})^{q-1}p_i' + T \tag{13}$$

where T contains terms involving $\bar{X}$ and higher powers of $\bar{X}$.

From (13) we obtain

$$\sigma^2_{\mu_q} = \sigma^2_{\mu_q'} + q^2\mu_{q-1}'^2\sigma^2_{\bar{X}} - 2q\mu_{q-1}'\sigma_{\bar{X}}\sigma_{\mu_q'}r_{\bar{X}\mu_q'} + U \tag{14}$$

where U involves $\bar{X}$ and higher powers. From (3), (8), (10) and (14) we have

$$\begin{aligned}\sigma^2_{\mu_q} &= \Sigma b^2_{q,i}\sigma^2_{p_i} + q^2\mu_{q-1}'^2\sigma^2_{p_i} - 2q\mu_{q-1}'\Sigma b_{q,i}\sigma^2_{p_i} + U \\ &= \Sigma(b_{q,i} - q\mu_{q-1}')^2\sigma^2_{p_i} + U.\end{aligned} \tag{15}$$

We now shift the origin to $\bar{X}$. All the terms in U vanish, μ_{q-1}' becomes μ_{q-1} , and the $b_{q,i}$ values go into $\beta_{q,i}$, where

$$\beta_{q,i} = (i - \bar{X})^q - (i - 1 - \bar{X})^q.$$

That is, (15) becomes

$$\sigma^2_{\mu_q} = \Sigma(\beta_{q,i} - q\mu_{q-1})^2\sigma^2_{p_i} . \tag{16}$$

It is of interest to give an alternative proof of the relation (16) possessing the desirable property of being very short and simple and at the same time yielding an expression for the $\beta_{q,i}$ in terms of $b_{r,i}(1 \leq r \leq q)$ and powers of $\bar{X}$.

If $x_1 + .5$ is taken as the origin, then (11) may be written as (17)

$$\mu_q = \Sigma(i - \bar{X})^q p'_i . \tag{17}$$

The application of the δ-operation to both sides of (17) yields:

$$\delta\mu_q = \Sigma(i - \bar{X})^q\delta p'_i - q\Sigma(i - \bar{X})^{q-1}p'_i\delta\bar{X} \tag{18}$$

$$= \Sigma\,(\beta_{q,i} - q\mu_{q-1})\delta p_i . \tag{19}$$

Repetition of a previous argument gives the result:

$$\sigma^2_{\mu_q} = \Sigma(\beta_{q,i} - q\mu_{q-1})^2\sigma^2_{p_i} . \tag{20}$$

In order to derive the relation connecting the $\beta_{q,i}$ with $b_{r,i}(1 \leq r \leq q)$ we expand $\Sigma(i - \bar{X})^q\delta p'_i$ in equation (18). This expansion yields:

$$\begin{aligned} \sum(i - \bar{X})^q\delta p'_i &= \sum(i^q - {}_qC_i\, i^{q-1}\bar{X} + {}_qC_2\, i^{q-2}\bar{X}^2 \\ &\quad + \cdots + (-1)^{q-1}\,{}_qC_{q-1}\, i\bar{X}^{q-1} + (-1)^q\bar{X}^q)\delta p'_i \\ &= \sum\,(b_{q,i} - {}_qC_1\, b_{q-1,i}\bar{X} + {}_qC_2\, b_{q-2,i}\bar{X}^2 \\ &\quad + \cdots (-1)^{q-1}q\bar{X}^{q-1}b_{1,i})\delta p_i \end{aligned} \tag{21}$$

i.e.,

$$\begin{aligned} \beta_{q,i} = b_{q,i} - {}_qC_1\, b_{q-1,i}\bar{X}\,{}_qC_2\, b_{q-2,i}\bar{X}^2 \\ + \cdots + (-1)^{q-1}q\bar{X}^{q-1}b_{1,i} . \end{aligned} \tag{22}$$

The relationship (16) combined with (22) enables one to compute the standard errors of a number of useful statistics. In particular in case $q = 2$ it follows that

$$\sigma^2_{\mu_2} = \sigma^2_{\sigma^2} = \Sigma(b_{2,i} - 2\bar{X})^2\sigma^2_{p_i} . \tag{23}$$

Combining (23) with the well-known result that

$$\sigma_\sigma = \sigma_{\mu_2}/2\sigma \tag{24}$$

we see that

$$\sigma_\sigma = \frac{\sqrt{\sum\{(2i - 3) - 2\bar{X}\}^2\sigma^2_{p_i}}}{2\sqrt{\sum\,(2i - 3)p_i - (\sum p_i)^2}} . \tag{25}$$

Formula (25) is useful in significance tests involving the standard deviations of sensitivity data.

3. Standard errors of the moments in standard units. We now turn our attention to the derivation of the standard error of the higher moments when expressed in standard units. Before proceeding with the derivation it is convenient to find the correlation between the qth and rth moments about the mean. This result is an immediate consequence of (19), for since

$$\delta\mu_q \doteq \Sigma\{\beta_{q,i} - q\mu_{q-1}\}\delta p_i$$

and

$$\delta\mu_r = \Sigma\{\beta_{r,i} - r\mu_{r-1}\}\delta p_i \tag{26}$$

it follows that

$$\delta\mu_q\delta\mu_r = \Sigma\{\beta_{q,i} - q\mu_{q-1}\}\{\beta_{r,i} - r\mu_{r-1}\}(\delta p_i)^2 + Z \tag{27}$$

where Z contains terms $\delta p_i\delta p_j (i \neq j)$. Hence, as before,

$$\sigma_{\mu_q}\sigma_{\mu_r}r_{\mu_q\mu_r} = \Sigma\{\beta_{q,i} - q\mu_{q-1}\}\{\beta_{r,i} - r_{\mu_{r-1}}\}\sigma^2_{p_i} \,. \tag{28}$$

Let us now derive the standard errors of the moments in standard units, i.e., of

$$\alpha_q = \mu_q/\sigma^q. \tag{29}$$

Now in general,

$$\delta\alpha_q = \frac{\sigma^q\delta\mu_q - q\sigma^{q-1}\mu_q\delta\sigma}{\sigma^{2q}} = \frac{\sigma\delta\mu_q - q\mu_q\delta\sigma}{\sigma^{q+1}} \tag{30}$$

and since

$$\delta\mu_2 = 2\sigma\delta\sigma\,, \quad \text{or} \quad \delta\sigma = \delta\mu_2/2\sigma \tag{31}$$

we have

$$\delta\alpha_q = \frac{2\sigma^2\delta\mu_q - q\mu_q\delta\mu_2}{2\sigma^{q+2}} \tag{32}$$

and hence

$$(\delta\alpha_q)^2 = \frac{4\sigma^4(\delta\mu_q)^2 + q^2\mu_q^2(\delta\mu_2)^2 - 4q\sigma^2\mu_q\delta\mu_q\delta\mu_2}{4\sigma^{2(q+2)}} \tag{33}$$

$$\sigma^2_{\alpha_q} = \frac{4\mu_2^2\sigma^2_{\mu_q} + q^2\mu_q^2\sigma^2_{\mu_2} - 4q\mu_2\mu_q\sigma_{\mu_q}\sigma_{\mu_2}r_{\mu_q\mu_2}}{4\mu_2^{q+2}}\,. \tag{34}$$

In this case, it follows that

$$\sigma^2_{\alpha_q} = \frac{\begin{array}{r}4\mu_2^2\sum(\beta_{q,i} - q\mu_{q-1})^2\sigma^2_{p_i} + q^2\mu_q^2\sum\beta^2_{2,i}\sigma^2_{p_i} \\ - 4q\mu_q\mu_2\sum(\beta_{q,i} - q\mu_{q-1})\beta_{2,i}\sigma^2_{p_i}\end{array}}{4\mu_2^{q+2}} \tag{35}$$

or

$$\sigma^2_{\alpha_q} = \frac{\sum(2\mu_2(\beta_{q,i} - q\mu_{q-1}) - q\mu_q\beta_{2,i})^2\sigma^2_{p_i}}{4\mu_2^{q+2}} \tag{36}$$

If the qth moment about the mean vanishes, then

$$(37)\qquad \sigma^2_{\alpha_q} = \frac{4\mu_2^2 \sum (\beta_{q,i} - q\mu_{q-1})^2 \sigma^2_{p_i}}{4\mu_2^{q+2}} = \frac{\sigma^2_{\mu_q}}{\mu_2^q}.$$

It is readily seen that the standard errors of the skewness and flatness are special cases of formula (36) when $q = 3$ and $q = 4$ respectively.

4. Some minimization problems. In the analysis of sensitivity data it is most desirable to minimize $\sigma^2_{\bar{X}}$ or $\sigma^2_{\sigma^2}$ in order to increase the precision of significance tests involving $\bar{X}$ or σ respectively. Therefore, it is of interest to solve the following problem: Suppose that we have a sample of size N which is to be subdivided into n samples of size N_i to be tested at a number of fixed levels $\{x_i\}$, $i = 1, 2 \cdots n$, $\sum_{i=1}^{n} N_i = N$. Then what choice of values $\{N_i\}$ will minimize

$$\sigma^2_{\bar{X}} = \sum_{i=1}^{n} \frac{p_i q_i}{N_i}, \qquad \text{where} \quad \sum_{i=1}^{n} N_i = N?$$

In order to solve this problem most quickly we use the method of Lagrange multipliers, i.e., we minimize the expression

$$(38)\qquad L_1(N_i, \lambda) = \sum_{i=1}^{n} \frac{p_i q_i}{N_i} + \lambda\left(\sum_{i=1}^{n} N_i - N\right).$$

Taking the partial derivatives with respect to N_i we obtain the n equations

$$(39)\qquad \frac{p_i q_i}{N_i^2} = \lambda, \qquad \text{i.e.,} \quad N_i = \frac{\sqrt{p_i q_i}}{\lambda^{1/2}}, \qquad i = 1, 2 \cdots, n.$$

Summing over all values of i we obtain

$$(40)\qquad N = \sum_{i=1}^{n} \frac{\sqrt{p_i q_i}}{\lambda^{1/2}} \quad \text{or} \quad \lambda^{1/2} = \frac{N}{\sum_{i=1}^{n} \sqrt{p_i q_i}};$$

i.e., the best choice of values for $\{N_i\}$ is given by

$$(41)\qquad N_i = \frac{N\sqrt{p_i q_i}}{\sum_{i=1}^{n} \sqrt{p_i q_i}}.$$

The value of $\sigma^2_{\bar{X}}$ for this choice of the set $\{N_i\}$ is

$$(42)\qquad \frac{\left(\sum_{i=1}^{n} \sqrt{p_i q_i}\right)^2}{N}.$$

It is obvious that this is actually a minimum. In particular, it is less than the value of $\sigma^2_{\bar{X}}$ for $N_i = N_j = N/n$ (the number of groups is n). This follows from the application of Schwartz' inequality to (42), for

$$(43)\qquad \frac{\left(\sum_{i=1}^{n} \sqrt{p_i q_i}\right)^2}{N} \leq \frac{n \sum_{i=1}^{n} p_i q_i}{N},$$

which equals the value of $\sigma^2_{\bar{X}}$ for $N_1 = N_2 = \cdots = N_n = N/n$. The equality holds if and only if $p_1 = p_2 = \cdots = p_n$.

Suppose next that we wish to minimize

$$\sigma^2_{\sigma^2} = \sum_{i=1}^{n} \frac{(b_{2,i} - 2\bar{X})^2 p_i q_i}{N'_i} = \sum_{i=1}^{n} \frac{\beta^2_{2,i} p_i q_i}{N'_i}, \tag{44}$$

where

$$\beta_{2,i} = b_{2,i} - 2\bar{X}.$$

We proceed as before to minimize the expression

$$L_2(N'_i, \lambda) = \sum_{i=1}^{n} \frac{\beta^2_{2,i} p_i q_i}{N'_i} + \lambda\left(\sum_{i=1}^{n} N'_i - N\right). \tag{45}$$

Taking partial derivatives with respect to N'_i we obtain

$$\frac{\beta^2_{2,i} p_i q_i}{N'^2_i} = \lambda \quad \text{i.e.} \quad N'_i = \frac{|\beta_{2,i}| \sqrt{p_i q_i}}{\lambda^{1/2}}, \qquad i = 1, 2, \cdots, n \tag{46}$$

or summing over all values of i we obtain

$$N = \sum_{i=1}^{n} \frac{|\beta_{2,i}| \sqrt{p_i q_i}}{\lambda^{1/2}} \quad \text{or} \quad \lambda^{1/2} = \sum_{i=1}^{n} \frac{|\beta_{2,i}| \sqrt{p_i q_i}}{N}; \tag{47}$$

i.e., the best choice of values for $\{N'_i\}$ is given by

$$N'_i = \frac{N |\beta_{2,i}| \sqrt{p_i q_i}}{\sum_{i=1}^{n} |\beta_{2,i}| \sqrt{p_i q_i}}. \tag{48}$$

The minimum value of $\sigma^2_{\sigma^2}$ is given by

$$\frac{\left(\sum_{i=1}^{n} |\beta_{2,i}| \sqrt{p_i q_i}\right)^2}{N}. \tag{49}$$

In practice we desire a set $\{N_i\}$ which will make $\sigma^2_{\bar{X}}$ and $\sigma^2_{\sigma^2}$ small simultaneously. Unfortunately this is not in general possible. In fact, it may be asserted that the set $\{N_i\}$ minimizing, $\sigma^2_{\bar{X}}$ will yield a large value of $\sigma^2_{\sigma^2}$ and similarly the set $\{N'_i\}$ minimizing $\sigma^2_{\sigma^2}$ will yield a large value of $\sigma^2_{\bar{X}}$. The reason for this curious behavior lies in the fact that the only difference between the set $\{N_i\}$ and the set $\{N'_i\}$ is the set of numbers $\{|\beta_{2,i}|\} = \{|(2i - 3) - 2\bar{X}|\}$. These numbers, however, change the character of the sets $\{N_i\}$ and $\{N'_i\}$. In particular $\{N'_i\}$ takes on its largest values for both small and large values of i, whereas $\{N_i\}$ takes on small values in these regions; $\{N'_i\}$ takes on small values for those values of i which are the integral values closest to $\bar{X} + 3/2$, whereas $\{N_i\}$ takes on large values for such values of i. It is this curious juxtaposition of $\{N_i\}$ and $\{N'_i\}$ that renders it impossible to choose sets of numbers $\{N_i\}$ minimizing $\sigma^2_{\bar{X}}$ and $\sigma^2_{\sigma^2}$ simultaneously.

NOTES

This section is devoted to brief research and expository articles, notes on methodology and other short items.

NOTE ON RUNS OF CONSECUTIVE ELEMENTS

By J. Wolfowitz

Columbia University

In my paper [1] I did not derive the asymptotic distribution of $W(R)$, an omission which I wish to correct in this note.

Let the stochastic variable $R = (x_1, \cdots, x_n)$ be a permutation of the first n positive integers, where each permutation has the same probability $\frac{1}{n!}$. A subsequence $x_{i+1}, x_{i+2}, \cdots, x_{i+l}$, is called a run of consecutive elements of length l if:

a) when l' is any integer such that $1 \leq l' < l$,

$$| x_{i+l'} - x_{i+l'+1} | = 1$$

b) when $i > 0, | x_i - x_{i+1} | > 1$

c) when $i + l < n, | x_{i+l} - x_{i+l+1} | > 1$.

Let $W(R)$ be the total number of runs in R. Then $n - W(R)$ is a stochastic variable which, it will be shown, has in the limit the Poisson distribution with mean value 2. More precisely, if $p(w)$ is the probability that $n - W(R) = w$, then

$$\lim_{n \to \infty} p(w) = \frac{2^w}{e^2 \cdot w!}. \tag{1}$$

Proof: Define stochastic variables $y_i (i = 1, 2, \cdots, n)$, as follows: $y_i = 1$ if x_i is the first element of a run of length 2, $y_i = 0$ otherwise. It is easy to see that the probability that $x_i (i = 1, 2, \cdots, n)$ be the initial element of a run of length greater than two is $O\left(\frac{1}{n^2}\right)$ and hence that the probability of the occurrence of a run of length greater than two is $O\left(\frac{1}{n}\right)$. Hence the limiting distribution of $n - W(R)$ is the same as that of

$$y = \sum_{i=1}^{n} y_i ,$$

provided either exists.

The y_i are dependent stochastic variables and almost all (i.e., all with the exception of a fixed number) have the same marginal distribution. We now wish to consider the expression

$$E(y_{i_1}^{\alpha_1} y_{i_2}^{\alpha_2} \cdots y_{i_k}^{\alpha_k})$$

(where the symbol E denotes the expectation) for any set of fixed positive integers $k, \alpha_1, \cdots, \alpha_k$, and for all k-tuples $i_1, i_2, \cdots, i_k$, with no two elements equal. Now

$$E(y_{i_1}^{\alpha_1} y_{i_2}^{\alpha_2} \cdots y_{i_k}^{\alpha_k}) = E(y_{i_1} y_{i_2} \cdots y_{i_k})$$

is the probability that $y_{i_1} = 1$, $y_{i_2} = 1, \cdots, y_{i_k} = 1$, simultaneously. This probability is either zero (for example, when $| i_2 - i_1 | = 1, | i_3 - i_2 | = 1$, etc. or when $i_1 = n$, etc.) or $\left(\frac{2}{n}\right)^k + O\left(\frac{1}{n^{k+1}}\right)$. Moreover, the ratio of the number of k-tuples $i_1, i_2, \cdots, i_k$ for which the probability is zero to the number of k-tuples for which the probability is $\left(\frac{2}{n}\right)^k + O\left(\frac{1}{n^{k+1}}\right)$ is $O\left(\frac{1}{n}\right)$. Let $Z_i (i = 1, \cdots, n)$ be independent stochastic variables each with the same distribution such that the probability that $Z_i = 1$ is $2/n$ and the probability that $Z_i = 0$ is $(n - 2)/n$. It follows readily that the limit, as $n \to \infty$, of the jth moment ($j = 1, 2, \cdots$, ad inf.) of y about the origin, is the same as the limit of the same moment of Z, where

$$Z = \sum_{i=1}^{n} Z_i .$$

Since the Z_i are independently distributed, and since each can take only the values 0 and 1, the probability of the value 1 being $2/n$, the jth moment of Z about the origin approaches, as $n \to \infty$,

$$\mu_j = e^{-2} \sum_{i=1}^{\infty} \frac{i^j 2^i}{i!},$$

which is the jth moment about the origin of the Poisson distribution with mean value 2. By the preceding paragraph, μ_j is also the limit of the jth moment of y about the origin. Now von Mises [2] has proved that if the jth moment ($j = 1, 2, \cdots$, ad inf.) of a chance variable X_n, ($n = 1, 2, \cdots$, ad inf.), approaches, as $n \to \infty$, the jth moment of a Poisson distribution, then the distribution of X_n approaches the Poisson distribution with corresponding mean value. From this it follows that y has in the limit the distribution (1). We have already shown that y and $n - W(R)$ have the same limiting distribution, so that the required result follows.

REFERENCES

[1] J. Wolfowitz, *Annals of Math. Stat.*, Vol. 13 (1942), p. 247.

[2] R. v. Mises, *Zeitschrift für die angewandte Math. und Mechanik*, Vol. 1 (1921), p. 298.

NOTE ON CONSISTENCY OF A PROPOSED TEST FOR THE PROBLEM OF TWO SAMPLES

By Albert H. Bowker

Columbia University

Certain tests for the hypothesis that two samples are from the same population assume nothing about the distribution function except that it is continuous. Since the power functions of these tests have not been obtained, optimum tests are not known. However, one desirable[1] test property, that of "consistency," has been introduced by Wald and Wolfowitz [1]. A test is called consistent if the probability of rejecting the null hypothesis when it is false (the power of the test) approaches one as the sample number approaches infinity. This is a logical extension of the familiar idea of consistency introduced by Fisher. It will be shown that a test recently proposed by Mathisen [2] is not consistent with respect to certain alternatives.

The test proposed by Mathisen [2] may be described briefly as follows: Given two samples, observe the number (m) of elements of the second sample whose values are less than the median of the first sample. The distribution of m is independent of the population distribution under the null hypothesis. Let $P\{m < a\}$ denote the probability of the relation in braces under the null hypothesis. If m_1 and m_2 are significance points $(m_1 > m_2)$ such that

$$\begin{aligned} P\{m > m_1\} &= \beta_1 \\ P\{m < m_2\} &= \beta_2 \\ \beta_1 + \beta_2 &= \beta < 1, \end{aligned} \tag{1}$$

the statistic m can be used to test the hypothesis at the significance level β. This is called the case of two intervals. The method is extended by using the two quartiles and the median of the first sample to define four intervals into which the elements of the second sample may fall. If the second sample is of size $4n$ and the number which actually falls in each interval is n_1, n_2, n_3, and n_4 respectively, the distribution of

$$C = \frac{\sum_{i=1}^{4} (n_i - n)^2}{9n^2} \tag{2}$$

is also independent of the population distribution under the null hypothesis. Then if C^* is a significance point, such that

$$P\{C > C^*\} = \beta' < 1, \tag{3}$$

C can be used as a test of the hypothesis at the level β'.

To show that Mathisen's test is not consistent, we shall consider first the case of two intervals. Let X and Y be two independent stochastic variables whose cumulative distribution functions $F(x)$ and $G(x)$ are continuous. Let $x_1 < x_2 \cdots < x_{2n+1}$ and $y_1 < y_2 \cdots < y_{2n}$ be sets of ordered independent observations on X and Y. Then m is such that

$$y_m < x_{n+1} < y_{m+1}.$$

Let m_1 and m_2 be the significance points of the distribution of m, defined by (1). Clearly m_1 and m_2 depend on n. We shall prove that the sequence

$$\frac{m_1(n)}{2n} \qquad n = 1, 2, \cdots \tag{4}$$

[1] For large samples.

converges to $\frac{1}{2}$. Since (4) is bounded, it has at least one limit point. Let h be such a limit point. If $h < \frac{1}{2}$ and $\frac{1}{2} - h = 3\delta$, then there exists a monotonically increasing subsequence of the integers $n_1, n_2, \cdots$ and a number N such that for $n_i > N$

$$\left| \frac{m_1(n_i)}{2n_i} - h \right| < \delta. \tag{5}$$

Clearly $m/2n$ converges stochastically to $\frac{1}{2}$. Hence if $0 < \epsilon < 1$ is any arbitrarily small number, we can select n so large that the probability is at least $1 - \epsilon$ that

$$\left| \frac{m}{2n} - \frac{1}{2} \right| < \delta. \tag{6}$$

Hence for n sufficiently large, $P\{m > m_1\}$ is at least $1 - \epsilon$, a contradiction with (1). A similar contradiction appears if $h > \frac{1}{2}$. Hence (4) has only one limit point, $\frac{1}{2}$. In the same way we can prove that the sequence

$$\frac{m_2(n)}{2n} \qquad n = 1, 2, \cdots. \tag{7}$$

also converges to $\frac{1}{2}$.

Let $0 < \delta \leq \frac{1}{6}$. Consider now two pairs of populations, A and B, described as follows:

$$\begin{array}{lll}
\text{A)} & F(x) \equiv G(x) \equiv x & (0 \leq x \leq 1) \\
\text{B)} & F(x) \equiv x & (0 \leq x \leq 1) \\
 & G(x) \equiv 0 & (0 \leq x \leq \frac{1}{2} - 2\delta) \\
 & G(x) \equiv (x - \frac{1}{2} + 2\delta)(\frac{1}{2} - \delta)/\delta & (\frac{1}{2} - 2\delta \leq x \leq \frac{1}{2} - \delta) \\
 & G(x) \equiv x & (\frac{1}{2} - \delta \leq x \leq \frac{1}{2} + \delta) \\
 & G(x) \equiv \frac{1}{2} + \delta & (\frac{1}{2} + \delta \leq x \leq 1 - \delta) \\
 & G(x) \equiv (\frac{1}{2} + \delta) + (x - 1 + \delta)(\frac{1}{2} - \delta)/\delta & (1 - \delta \leq x \leq 1)
\end{array}$$

For both A and B, $F(x) \equiv G(x) \equiv 0$ for $x < 0$ and $F(x) \equiv G(x) \equiv 1$ for $x > 1$. For B, it will be shown that there exist values of n greater than any preassigned arbitrarily large number, such that the probability of rejecting the hypothesis when it is false is less than $\beta_1 + \beta_2 + \epsilon$ where ϵ is an arbitrarily small positive number.

Let h_1, h_2, h_3 denote the number of observations on X which fall in the intervals $0 < x \leq \frac{1}{2} - \delta$, $\frac{1}{2} - \delta < x \leq \frac{1}{2} + \delta$, $\frac{1}{2} + \delta < x \leq 1$ respectively for a fixed value of n. Let h_1', h_2', h_3' be the corresponding numbers for Y. For a fixed n, the probability of a set $h_1, h_2, h_3, h_1', h_2', h_3'$ is the same whether the samples be drawn from A or B. From (4), (7), and the stochastic convergence of $m/2n$, it follows that we can find an N such that for all $n > N$ the probability is at least $1 - \epsilon/2$ of the occurrence of a set h_1, h_2, h_3, h_1', h_2', h_3' for which y_{m_1}, x_{n+1}, y_{m_2} will fall in the interval $(\frac{1}{2} - \delta, \frac{1}{2} + \delta)$. Furthermore, for fixed h_2, h_2' the distribution within the interval is the same whether the sample came from A or B. Hence, even when the sample is drawn from B, for n sufficiently large,

$$P\{m > m_1\} < \beta_1 + \frac{\epsilon}{2}$$

$$P\{m < m_2\} < \beta_2 + \frac{\epsilon}{2}.$$

That is, for samples of sufficiently large size from B, the probability of rejecting the null hypothesis is at most $\beta_1 + \beta_2 + \epsilon$. Since $\beta_1 + \beta_2 < 1$ and ϵ is arbitrarily small, the probability can be made less than one and the test is not consistent in the case of two intervals.

In the case of four intervals, the proof is similar. In this case, we assume that the second sample has size $4n$. Clearly, $n_1/4n$, $n_2/4n$, $n_3/4n$, and $n_4/4n$ converge stochastically to $\frac{1}{4}$. If C^* is the significance point defined by (3), the sequence

$$C^*(n) \qquad\qquad n = 1, 2, \cdots$$

converges to zero. Now consider two pairs, A and B, of populations. A is the same as before and B consists of one uniform distribution and one which is identical with the uniform distribution in small intervals containing $x = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and 1, but is different everywhere else. As before, $F(x) \equiv G(x) \equiv 0$ for $x < 0$ and $F(x) \equiv G(x) \equiv 1$ for $x > 1$. Then for B, when n is large, the behavior of C, except for a probability arbitrarily near zero, will depend only on the intervals of coincidence. Hence for B

$$P\{C > C^*\} < \beta' + \epsilon$$

where ϵ is any arbitrarily small positive quantity.

Returning to the case of two intervals, if the samples are from different populations and if their cumulative distribution functions are identical in the neighborhood of their medians, the test is not consistent. If such a possibility is excluded from the class of admissible alternatives, we may expect that the test will be consistent. For example, if the class of alternatives is limited to those where $G(x) \equiv F(x + c)$, c a constant, the test will be consistent. A similar remark holds for the case of four intervals or for any fixed finite number of intervals. It appears, however, that if the number of intervals is a function of the sample size (say $\sqrt{n}$) and becomes infinite with sample size, a test of this kind will be consistent with respect to a general class of alternatives.

REFERENCES

[1] A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," *Annals of Math. Stat.*, Vol. XI (1940), pp. 147-62.

[2] H. C. Mathisen, "A method of testing the hypothesis that two samples are from the same population," *Annals of Math. Stat.*, Vol. XIV (1943), pp. 188-94.

HENRY LEWIS RIETZ—IN MEMORIAM

By A. R. Crathorne

University of Illinois

Forty odd years ago few if any American college catalogs mentioned the words "mathematical statistics." The word "actuary" often called for the use of a dictionary. Some courses in the theory of probability, theory of errors, or method of least squares touched on some phases of statistics but aside from this there was little interest in the subject. In England at this time Karl Pearson was well started in his work at University College but "Student" was an undergraduate student. In Germany, Lexis was finishing his somewhat unrecognized labors at Goettingen. In Denmark, Thiele, and in Norway, Charlier were lecturing and writing on statistics from their own individual viewpoints.

During the four decades which have passed, the interest in theoretical statistics in the United States has increased to the point where it has a well established journal of its own and few university mathematical departments fail to list statistical courses. In this growth no one has had more influence than the subject of this memoir. His published papers, his personality, his students and his well directed energy have all been more than helpful in putting mathematical statistics where it is today.

Henry Lewis Rietz, son of Jacob and Tabitha Jane Rietz, was born August 24, 1875 at Gilmore, Ohio. He attended the local schools and in 1895 entered Ohio State University receiving his B.S. degree in 1899. After graduation he went to Cornell University as scholar, then fellow and assistant in mathematics. During his stay at Cornell he was closely associated with two other mathematical students, J. W. Young and H. W. Kuhn, later heads of the departments of mathematics at Dartmouth and Ohio State University respectively. In his last year Rietz was particularly interested in group theory and worked for his doctorate with Professor G. A. Miller who was then a member of Cornell's faculty. His dissertation was "On primitive groups of odd orders," published later in the *American Journal of Mathematics* and referred to in the *Encyclopedie des Sciences Mathematiques*. After receiving the Ph.D. in 1902 he spent one year as professor of mathematics and astronomy at Butler College in Indianapolis.

In 1903, Rietz accepted an instructorship at the University of Illinois where he stayed until 1918 becoming full professor in the meantime. In 1918 he was called to the University of Iowa as head of the department of mathematics, a position he held until his retirement in 1942.

During his first year at Illinois his interests were mainly in pure mathematics. His advanced courses were "Theory of Invariants" during the first semester and "Higher Plane Curves" during the second. During the next year a demand arose for some course in statistics. None of the members of the mathematics department were particularly prepared to give such a course but Rietz was induced to try it. The result was that he offered a course "Averages and

Mathematics of Investment." This curious title was evidence of the fact that actuarial science had not reached the independent position that it has at the present time. In the following year he was appointed to the position of statistician of the College of Agriculture and from that time on during his stay at Illinois he divided his time equally between the department of mathematics and that college. His work as statistician was mainly supervision of the statistical work in the published bulletins. The first publication in the statistical field under his name was the 32-page appendix to Dean Davenport's treatise on breeding.

The first published statistical study was in 1908, a master's thesis for Miss Shade on "Correlation of efficiency in mathematics and in other subjects," printed as one of a series of University Studies. It is interesting to recall the attention which this paper received, especially from educational circles. It seemed to fix the method and form of calculation of correlation coefficients which occupied the time of many people during the following years. In these early years it was rather difficult to find a place of publication for a mathematical paper on statistics. Mathematical journals were somewhat reluctant in accepting articles. I remember one occasion when he jokingly complained of the correspondence necessary to explain to an editor the word "correlation" used in a paper.

From 1908 on, Rietz published a long list of papers on statistical topics, some purely theoretical, some expositional, some arising out of his connection with the college of Agriculture. Together with his later actuarial studies the list totals 150 titles, the more important of which are included in this article. His much quoted paper of 1920 on "Urn Schemata" pleased him more than any other paper. His little book *Mathematical Statistics*, one of the Carus mathematical monographs, written in 1926 was the basis for many university courses for years afterward.

In 1909 the American Institute of Actuaries was organized in Chicago, and Rietz was a charter member. He took particular delight in this organization and was rarely absent from the meetings. He was elected vice-president in 1919. He liked meeting practical actuaries and had a wide acquaintance among them. In 1916 he was appointed a member of the Illinois Pension Laws Commission and became its actuary. From that time on his interests were pretty evenly divided between mathematical statistics and actuarial problems connected with pensions. He was appointed actuary of the Chicago Pension Commission in 1926, was consulting actuary for the Presidents National Committee on Economic Security 1934, and was a member of the board of trustees of the Teacher's Insurance and Annuity Association 1934–38. His services as a consulting actuary were sought by a great many pension projects both in educational and in business circles. When he went to the University of Iowa in 1918, he accented actuarial theory in his teaching. Under his leadership the department became an outstanding school in this field. Many of his students hold prominent positions in the actuarial world.

In 1923, Rietz with eight others was appointed a member of the Committee on the Mathematical Analysis of Statistics of the division of Physical Sciences of the National Research Council. The work of this committee developed into the preparation of the "Handbook of Mathematical Statistics" with Rietz as Editor-in-Chief. This work had considerable use for a number of years after its publication and is an important part of the history of mathematical statistics in this country. A Russian edition of this book appeared in 1927 with a very long preface as a sort of apology for the translation. A few excerpts from this preface are—"Mathematical statistics is a purely technical weapon, politically unbiased, can serve with equal facility either to thwart or to expedite the movement for the emancipation of the proletariat depending in whose hands it happens to be;" "Mathematical statistics has nothing to do with philosophical enlightenment;" "Hence this book harbors no dangers for a soviet reader;" "The fact that the western authors work in a bourgeois society has no bearing on their methods."

The Institute of Mathematical Statistics was organized in 1935 with Rietz as the informal chairman of a steering committee during the months of discussion preceding the organization. He thus became the logical first president. He has taken a more than active interest in the Institute,—as a contributor to the *Annals*, as one of its editors, as general counselor, as a good friend. In appreciation of this and in recognition of his contributions to the initiation and development of mathematical statistics in America, the 1943 volume of the *Annals of Mathematical Statistics* was dedicated to him, on the occasion of his retirement after twenty five years of service as head of the department of mathematics at the University of Iowa.

Professor Rietz received many honors in other fields. He was President of the Mathematical Association of America in 1924, vice-president of the American Statistical Association in 1925, vice-president of the American Mathematical Society 1928–9, and a member of the editorial staffs of the Bulletin and the Transactions of that society for many years, president of the Iowa Academy of Science, 1931. He was starred in American Men of Science, a fellow of the Royal Statistical Society of London, and of the American Association for the Advancement of Science. He took great interest in local affairs and held many offices in church, social and business organizations in Iowa City. His mind was not altogether centered upon research and the development of mathematics and statistics. He took great pride in his teaching. He was the principal author of a number of college texts in mathematics that had wide use, and was on many committees concerned with the problem of teaching mathematics to undergraduates.

At the time of his retirement, Professor Rietz was in failing health and was practically an invalid until the time of his death at the University Hospital at Iowa City on December 7, 1943. He leaves a brother, Professor John Rietz, Morgantown, West Virginia, and a sister, Mrs. T. S. Taylor, Caldwell, New Jersey.

Selected Titles of Publications by H. L. Rietz

1. "On primitive groups of odd order," *Amer. Jour. of Math.*, Vol. 26 (1904), pp. 1-30.
2. "On groups in which certain commutative operations are conjugate," *Trans. Amer. Math. Soc.*, Vol. 5(1904), pp. 500–508.
3. "Simply transitive groups which are simple groups," *Bulletin Amer. Math. Soc.*, Vol. 11(1905), pp. 545–46.
4. "Statistical Methods. Appendix to Principles of Breeding." *A Treatise on Thremmatology* by E. Davenport, Ginn and Co., Boston, 1907, pp. 681–713.
5. "Correlation of efficiency in mathematics and efficiency in other subjects—A Statistical Study," Rietz and Shade. The University Studies, Urbana, Illinois, November 1908, 20 pp.
6. *Statistical Methods Applied to the Study of Type and Variability in Corn*, Eugene Davenport and Henry L. Rietz, Bulletin No. 119, Illinois Agricultural Experiment Station, 1907.
7. "On inheritance in the production of butter fat," *Biometrika*, London, Vol. 7(1909), pp. 106–126.
8. "On a mean difference problem that occurs in statistics," *Amer. Math. Month.*, Vol. 17(1910), pp. 235–40.
9. "On the metabolism experiment as a statistical problem," Rietz and Mitchell. *Jour. of Biol. Chem.*, 1910.
10. *On the Measurement of Correlation with Special Reference to Some Characters of Indian Corn*, Henry L. Rietz and Louis H. Smith, Bulletin No. 148, University of Illinois Agricultural Experiment Station, November 1910.
11. "On the construction and graduation of a rural life table," Rietz, H. L. and Forsyth, C. H., *Record Amer. Inst. of Actuar.*, Vol. 1(1911), pp. 9–19.
12. "On the theory of correlation with special reference to certain significant loci on the plane of distribution in the case of normal correlation," *Annals of Math.*, Vol. 13(1912), pp. 187–199.
13. "Note on the definition of an asymptote," *Amer. Math. Month.*, Vol. 19 (1912), pp. 89–90.
14. "The determination of the relative volumes of the components of rocks by mensuration methods," Lincoln and Rietz. *Economic Geology*, Vol. VIII, No. 2, March, 1913.
15. "On the mathematical theory of risk and Landre's theory of the maximum," *Record Amer. Inst. Actuar.*, Vol. 11(1913), pp. 1–14.
16. "On the status of certain current pension funds," *Record Amer. Inst. Actuar.*, Vol. 3(1914), pp. 33–53.
17. "Group Insurance," *Record Amer. Inst. Actuar.*, Vol. 3(1914), pp. 277–79.
18. "Degrees of resemblance of parents and offspring with respect to birth as twins for registered Shropshire sheep." (Rietz, H. L. and Roberts, Elmer.) *Jour. Agric. Res.*, Vol. 4, No. 6, pp. 479-510.
19. "Note on double interpolation by finite differences," *Record Amer. Inst. Actuar.*, Vol. 4(1915), pp. 15–22.
20. "On the correlation of marks in mathematics and law," *Jour. Educ. Psych.*, Vol. 7, No. 2, pp. 87–92.
21. "The operation of pension laws in foreign countries," Report of Illinois Pension Laws Commission, 1916, pp. 19–35.
22. "Actuarial Report on Pension Funds for Public Employees of Illinois," H. L. Rietz, and D. F. Campbell, Report of Illinois Pension Laws Commission, 1916, pp. 72–199.
23. "The underlying principles of a pension plan," H. L. Rietz, G. E. Hooker, and D. F. Campbell, Report of Illinois Pension Laws Commission, 1916, pp. 272–284.
24. "Report of Illinois Pension Laws Commission," Rietz and others, 1916, pp. 310.

25. "On the value of certain proposed refunds payable at the death of an annuitant under a pension system," *Record Amer. Inst. Actuar.*, Vol. 6(1917), pp. 62–75.

26. "A statistical study of some indirect effects of certain selections in the breeding of indian corn," Rietz, H. L. and Smith, L. H., *Jour. of Agric. Res.*, Vol. 11(1917), pp. 105–46.

27. "Methods of providing for expenses of new business by life insurance companies," *Amer. Econ. Rev.*, Vol. 7(1917), pp. 832–38.

28. "A Report to the Trustees of the Carnegie Foundation by the Commission Chosen to study and report upon the Proposed Plan of Insurance and Annuities," Rietz and others, 19 pp.

29. "Pensions for public employees," *Amer. Polit. Sci. Review*, Vol. 12(1918), pp. 265–68.

30. "Statistical methods for preparation for war department service," *Amer. Math. Month.*, Vol. 26(1919), pp. 99–100.

31. "Recent developments in pension plans for public employees," *Record Amer. Inst. Actuar.*, Vol. 8(1919), pp. 1–12.

32. "Scope and advantages of courses of instruction on life insurance in American colleges and universities," *Record Amer. Inst. Actuar.*, Vol. 8(1919), pp. 202–06.

33. "The effect of present inflated prices on the future interest rate," *Record Amer. Inst. Actuar.*, Vol. 8(1919), pp. 308–14.

34. "On functional relations for which the coefficient of correlation is zero," *Quart. Pub. Amer. Stat. Assn.* September 1919, pp. 472–76.

35. "Exposition of the main provisions of the standard plan for a combined comprehensive annuity and insurance system for public employees," Illinois Pension Laws Commission Report, 1919, pp. 25–48.

36. "Recent developments in pension legislation in other states of the United States," Illinois Pension Laws Commission Report, 1919, pp. 201–06.

37. "The world's experience in the operation of public service pension systems," Illinois Pension Commission Report, 1919, p. 207–14.

38. "Summary of the report of the Illinois Pension Laws Commission of 1916," Illinois Pension Commission Report, 1919, pp. 215–22.

39. "Effects of the pension legislation by the fiftieth general assembly of Illinois," Illinois Pension Commission Report, 1919, pp. 223–34.

40. "Illinois state teachers' pension and retirement system. Illinois Pension Commission Report," 1919, pp. 235–38.

41. "Industrial and institutional pension systems," Illinois Pension Commission Report, 1919, pp. 239–50.

42. "Urn schemata as a basis for the development of correlation theory," *Annals of Math.*, Vol. 21(1920), pp. 306–22.

43. "On certain properties of Makeham's laws of mortality," *Amer. Math. Month.*, Vol. 27(1921), pp. 152–65.

44. "Pension systems for insurance company employees," *Record Amer. Inst. Actuar.*, Vol. 10(1921), pp. 1–14.

45. "An elementary exposition of the theorem of Bernoulli with applications to statistics," *Math. Teacher*, Vol. 14(1921), pp. 427–34.

46. "Frequency distributions obtained by certain transformations of normally distributed variates," *Annals Math.*, Vol. 22(1922), pp. 292–300.

47. "'Statistics' in a Mathematical Encyclopedic Dictionary," *Amer. Math. Month.*, Vol. 29(1922), pp. 333–337.

48. "On the subject matter of a course in mathematical statistics." Presented before the Mathematical Association of America at the Symposium held in Cambridge, Dec. 29, 1922, *Amer. Math. Month.*, Vol. 30(1923), pp. 155–166.

49. "On certain topics in the mathematical theory of statistics," symposium Lectures before American Mathematical Society, *Bull. Amer. Math. Soc.*, Vol. 30(1924), pp. 417–453.

50. "On annuity rates," *Record Amer. Inst. Actuar.*, Vol. 13(1924), pp. 120–122.
51. "On certain applications of mathematical statistics to actuarial data," *Record Amer. Inst. Actuar.*, Vol. 13(1924), pp. 214–250.
52. "On a certain law of probability of Laplace," *Proc. Internat. Math. Cong.*, Toronto, Vol. 2(1924), pp. 795–799.
53. "Note on average numbers of brothers and of sisters of the boys in families of N children," *Science*, Vol. 60(1924), pp. 46–47.
54. "On the representation of a certain fundamental law of probability," *Trans. Amer. Math. Soc.*, Vol. 27(1925), pp. 197–212.
55. "On applications of statistical methods in actuarial science," *Record Amer. Inst. Actuar.*, Vol. 14(1925), pp. 102–04.
56. "Mathematical background for the study of statistics," H. L. Rietz and A. R. Crathorne, *Jour. Amer. Stat. Assn.*, Vol. 21(1926), pp. 435–440.
57. "On certain applications of the differential and integral calculus in actuarial science," *Amer. Math. Month.*, Vol. 33(1926), pp. 9–23.
58. "Note on the most probable number of deaths," *Record Amer. Inst. Actuar.*, Vol. 16(1927), pp. 25–29.
59. "On certain properties of frequency distributions of the powers and roots of the variates of a given distribution," *Proc. Nat. Acad. Sci.*, Vol. 13(1927), pp. 817–20.
60. Discussion of "Interpolation with modified coefficients," *Record Amer. Inst. Actuar.*, Vol. 16(1927), pp. 232–33.
61. "On the risk problem from a mathematical point of view," *Neuvienne Congres International D'Actuaires, Rapports, Tome II*, (1930), pp. 294–306.
62. "Pensions for superannuated employees," Retiring vice-presidential address before Section K. AAAS, *Sci. Month.*, March 1930, pp. 224–30.
63. "On certain properties of frequency distributions obtained by a linear fractional transformation of the variates of a given distribution," *Annals of Math. Stat.*, Vol. 2(1931), pp. 38–47.
64. "Note on the distribution of the standard deviation of sets of three variates drawn at random from a rectangular distribution," *Biometrika*, Vol. 23(1931), pp. 424–26.
65. "Some remarks on mathematical statistics," Retiring president's address before the Iowa Academy of Science, *Science*, Vol. 74(1931), pp. 1–4.
66. "Comments on applications of recently developed theory of small samples," *Jour. Amer. Stat. Assn.*, Vol. 26(1931), pp. 37–44.
67. "A simple non-normal correlation surface," *Biometrika*, Vol. 24(1932), pp. 288–90.
68. "On the Lexis theory and the analysis of variance," *Bull. Amer. Math. Soc.*, Vol. 27(1932), pp. 731–35.
69. "Unemployment and social insurance," *Record Amer. Inst. Actuar.*, Vol. 23(1934), pp. 147–52.
70. "On the frequency distribution of certain ratios," *Annals of Math. Stat.*, Vol. 7(1936), pp. 145–53.
71. "Some topics in sampling theory," *Bull. Amer. Math. Soc.*, Vol. 43(1937), pp. 209–30.
72. "Collective insurance," *Bull. Amer. Assn. Univ. Prof.*, Vol. 23(1937), pp. 278–81.
73. "On the distribution of the 'Student' ratio for small samples from certain non-normal distributions," *Annals Math. Stat.*, Vol. 10(1939), pp. 265–74.
74. "On a recent advance in statistical inference," *Amer. Math. Month.*, Vol. 45(1938), pp. 149–58.

Doctorate Dissertations Written Under the Supervision of Professor Rietz

Reilly, John Franklin, 1921. On certain generalizations of osculatory interpolation.

Weida, Frank M., 1923. The valuation of life annuities with refund of an arbitrarily assigned part of the purchase price.

Smith, Clarence De Witt, 1928. On generalized Tchebycheff inequalities in mathematical statistics.

Meyer, Herbert A., 1929. On certain inequalities with applications in actuarial science.

Craig, Allen Thornton, 1931. On the distribution of certain statistics derived from small random samples.

Wilks, Samuel Stanley, 1931. On the distribution of statistics in samples from a normal population of two variables with matched sampling of one variable.

Fischer, Carl H., 1932. On correlation surfaces of sums with a certain number of random elements in common.

Harper, Floyd S., 1935. An actuarial study of infant mortality.

Olliver, Arthur, 1935. On certain mathematical developments underlying an analysis of general death rates.

Knowler, Lloyd A., 1937. Actuarial aspects of recent old age security legislation.

Olshen, Abraham C., 1937. Transformations of the Pearson type III distribution.

Berg, William D., 1941. Theorems on certain type A difference equation graduations.

Satterthwaite, Franklin, 1941. Developments on the theory of Chi-square.

Garfin, Louis, 1942. A comparative study of the underlying principles of certain pension schemes for a staff of employees with special reference to teachers and public employees.

'The last two dissertations were under the joint supervision of Professor Rietz and Professor A. T. Craig.

Books

1. *College Algebra*, H. L. Rietz and A. R. Crathorne, First edition, 1909, Fourth edition, 1939. Henry Holt and Company, New York.
2. *School Algebra*, two volumes. H. L. Rietz, A. R. Crathorne, E. H. Taylor, 1915. Henry Holt and Company.
3. *Mathematics of Finance*, H. L. Rietz, A. R. Crathorne, J. C. Rietz. First edition 1921, second edition 1929. Henry Holt and Company.
4. *Introductory College Algebra*. H. L. Rietz and A. R. Crathorne. First edition 1923, second edition 1933. Henry Holt and Company.
5. *Handbook of Mathematical Statistics*, H. L. Rietz, Editor-in-chief, 1924, Houghton Mifflin Company.
6. *Mathematical Statistics*. Third Carus Mathematical Monograph, published for the Mathematical Association of America by the Open Court Publishing Co. 1927.
7. *Review of Pre-college Mathematics*, C. J. Lapp, F. B. Knight, H. L. Rietz, 1934. Scott, Foresman and Company.
8. *Plane Trigonometry*, H. L. Rietz, J. F. Reilly, Roscoe Woods, 1935. The Macmillan Company.
9. *Plane and Spherical Trigonometry*. H. L. Rietz, J. F. Reilly, Roscoe Woods, 1936. The Macmillan Company.
10. *Intermediate Algebra*, H. L. Rietz, A. R. Crathorne, L. J. Adams, 1942. Henry Holt and Company.
11. *Review of Mathematics for College Students*. C. J. Lapp, F. B. Knight, H. L. Rietz, 1942. Foresman and Company.

NEWS AND NOTICES

Readers are invited to submit to the Secretary of the Institute news items of general interest

Personal Items

Lt. Col. Joseph Berkson is now stationed at Headquarters, Army Air Forces, Air Surgeon's Office, Washington 25, D. C.

Dr. Ernest E. Blanche is now Statistical Director, Office of the Director of Engineering, with the Curtiss-Wright Corporation at Buffalo.

Dr. Alva E. Brandt is overseas serving as Operations Analyst for the Army Air Forces.

Professor W. G. Cochran and Dr. A. M. Mood are serving as Research Mathematicians on a war research project at Princeton University. Professor Cochran is on leave of absence from Iowa State College; Dr. Mood from the University of Texas.

Mr. Robert Dorfman is overseas serving as Operations Analyst with the 13th U. S. Air Force.

Mr. R. M. Foster of Bell Telephone Laboratories has been appointed professor and head of the department of mathematics of the Polytechnic Institute of Brooklyn.

Dr. Andrew I. Peterson is now Director of Manufacturing Research at the Victor Division of the Radio Corporation of America, Camden, N. J.

Dr. Edward Helly, visiting lecturer at the Illinois Institute of Technology, died on November 28, 1943.

Professor Henry L. Rietz died on December 7, 1943 after a long period of illness. An account of Professor Rietz' scientific life and achievements by Professor A. R. Crathorne appears on pp. 102–108 of the present issue of the *Annals*.

New Members

The following persons have been elected to membership in the Institute:

Allen, Roy George Douglas. D.Sc. (London). Reader in Economic Statistics, University of London. Apt. 219, 2745 29th St. NW, Washington 8, D. C.

Burk, Mrs. Marjorie F. A.B. (Hunter) Assoc. Statistician, teorology, HQ, AAF. 1912 Third St. NE, Washington 2, D. C.

Churchman, C. West. Ph.D (Pennsylvania) Research Statistic . Frankford Arsenal; and Lecturer in Philosophy, Univ. of Pennsylvania. Frankfora Arsenal, Philadelphia, Pa.

Crump, S. Lee. B.S. (Cornell) Research Associate, Iowa State College. Statistical Lab , Iowa State College, Ames, Iowa.

Divatia, M. V. M.A. (Columbia) Statistical Officer, Dept. of Industries and Civil Supplies, New Delhi, India.

Freund, John E. B.A. (U.C.L.A.) Box 4221 Westwood Village Station, Los Angeles 24, Calif.

Gill, John P. Statistician, Dept. of Research and Statistics, Federal Reserve Bank, Dallas, Texas.

Lindsey, Fred D. M.A. (George Washington Univ.) 628 West 114 St., New York, N. Y.

Maloney, Clifford J. M.A. (Minnesota) 2d Lt., Sig. C. 1535 18th St. N., Arlington, Va.

Mandel, John. Licence en Sciences (Univ. of Brussels) 45 Kew Gardens Rd., Kew Gardens, N. Y.

Mathieus, George John. B.L. (Univ. of Dayton) Shop followup—Assembly Planning, Douglas, Long Beach. 1853 Poppy St., N. Long Beach, Calif.

McIntyre, Francis E. Ph.D. (Chicago) Program Officer, Foreign Economic Administration. Rm. 2446 Temporary U Bldg., Washington 25, D. C.

Raybould, Ethel H. M.A. (Queensland) Lecturer in Mathematics. The University of Queensland, Brisbane, Australia.

Rosenblatt, Alfred. Ph.D. (Cracow) Catedratico de la Universidad de San Marcos, Lima. Calle Atahualpa 192, Miraflores, Peru.

Saunders, Robert J. B.S. (Mass. Inst. of Tech.) Captain, Ordnance Dept., Inspection Sec., Amm'n Branch, Office Chief of Ordnance. Office Field Director Ammunition Plants, 3637 Lindell Blvd., St. Louis 8, Missouri.

Schwartz, David H. B.S. (C.C.N.Y.) Assoc. Statistician, Office of the Quartermaster General. 338 N. Geo. Mason Dr., Arlington, Va.

Solomons, Leonard M. B.A. (Columbia) Time Study Man. 150-11 88 Ave., Jamaica 2, N. Y.

Steinberg, Joseph. B.S. (C.C.N.Y.) Associate Statistician, Bureau of Research and Statistics, Social Security Board. 5041 North Capitol Street, Washington 11, D. C.

Tomlinson, Malcolm C. W. 3820 Southern Ave., S.E., Washington, D. C.

Wilson, Edward F. Asst. Engineer, Special Projects. Bldg. 650, Research Center, Aberdeen Proving Ground, Md.

Announcements

Washington Meeting of the Institute

There will be a joint sectional meeting* of the Institute of Mathematical Statistics and the American Statistical Association at the Hotel Statler and George Washington University in Washington, D. C., on Saturday and Sunday, May 6-7, 1944.

On Saturday afternoon there will be a session on the *Theory of Statistical Inference* with Professor A. Wald and Lt. J. H. Curtiss as the speakers. On Sunday morning there will be a session on contributed papers and on Sunday afternoon a session will be devoted to *Time Series* with Professor J. L. Doob and Dr. T. Koopmans as the speakers.

Summer Meeting of the Institute

The summer meeting of the Institute will be held in conjunction with the summer meetings of the American Mathematical Society and the Mathematical Association of America, at Wellesley College, Wellesley, Massachusetts, on August 12-14, 1944. Abstracts of contributed papers for this meeting should be sent (in duplicate) to the Secretary of the Institute before July 1, 1944.

* The Washington meeting of the Institute had been announced in the December, 1943 issue of the *Annals* for April 27-28, 1944, but the date has been revised to May 6-7, in order that the meeting may be held jointly with the Association.

ANNUAL REPORT OF THE PRESIDENT OF THE INSTITUTE

During 1943, the second war year in which a regular annual meeting of the Institute was not held, the business of the Institute was necessarily largely conducted by mail. As reported in the March issue of the *Annals*, the Secretary, the Editor of the *Annals*, and I met in Pittsburgh in January for discussion of the Institute's affairs. Since three members of the Board of Directors do not constitute a quorum, certain proposals arising from this informal meeting were submitted by mail to the whole Board for action. At the fall meeting of the Institute held in New Brunswick a quorum of the Board was present, Hotelling, Wald, Wilks, Craig, and action was taken on certain matters. The chief subject for consideration was future meetings of the Institute. It was agreed that again in accordance with the request of the O.D.T. no annual meeting should be planned for 1943. Because of the success of local meetings held in New York City in May and in Washington in June, it was voted to repeat these in 1944. The New Brunswick meeting was also very successful, and it is hoped that we may again hold our fall meeting in conjunction with those of the American Mathematical Society and the Mathematical Association of America in 1944. The desirability of other local meetings in addition to those in New York and Washington was recognized but, so far, I know of no plans for any in 1944.

During the year two local chapters of the Institute were organized in Pittsburgh and Washington in accordance with regulations adopted by the Board of Directors, and these were recognized by the Board. The sponsor for the Pittsburgh group is E. G. Olds, and W. G. Madow has the same responsibility for the Washington chapter.

The Membership Committee for 1943 consisted of Vice-President Deming, Chairman, W. G. Cochran, P. S. Dwyer, and A. J. Lotka. As the result of their recommendation the following eighteen new Fellows of the Institute were elected by the Board of Directors: A. H. Copeland, J. H. Curtiss, J. F. Daly, C. E. Dieulefait, H. F. Dodge, Churchill Eisenhart, Will Feller, Milton Friedman, M. A. Girschick, M. H. Hansen, P. G. Hoel, Tjalling Koopmans, A. M. Mood, L. J. Reed, L. E. Simon, F. F. Stephan, W. R. Thompson, and Jacob Wolfowitz.

W. D. Baten, L. A. Aroian, I. W. Burr, and H. F. Dodge, at the request of the Board, continued to serve as a committee for securing additional library subscriptions to the *Annals*. Programs for the New York and New Brunswick meetings were in charge of Vice-President Wald, and that for the Washington meeting was arranged by W. G. Madow.

A proposal originating with Harold Hotelling that the Institute consider petitioning the Federal Government that the W. P. A. Computing Project be made a permanent computing group for the service of scientific research in the construction of important numerical tables, was referred to a committee consisting of A. R. Crathorne, Chairman, P. S. Dwyer, and Will Feller. As a result of their report, the Board appointed the following permanent committee on tabular computation, P. S. Dwyer, Chairman, Churchill Eisenhart, and Will Feller. It is expected that this committee will cooperate with representatives of other scientific organizations interested in tabular computation.

The Nominating Committee for the recently held election of the Institute was A. T. Craig, Chairman, B. H. Camp, and J. H. Curtiss. G. W. Snedecor served the Institute as its representative on the Council of the American Association for the Advancement of Science.

To all of those mentioned the Institute is indebted, and for the Board of Directors I wish to thank them for their special contributions to the Institute.

I have reserved for particular mention the services of Vice-President Deming who was appointed by the Board the official representative of the Institute to deal with officials of the Selective Service Board in Washington relative to the deferment from military service of competent and experienced statisticians engaged in work important for the prosecution of the war. He gave much time and effort to this on his own initiative, and the Institute and all statisticians owe him much in that he has been very successful in convincing important Washington officials of the value of the services being rendered by good statisticians.

In December the Board with deep regret accepted the resignation of E. G. Olds who was completing his third year as Secretary of the Institute. No member of the Institute will need to be reminded of the devotion and efficiency with which he filled his office. A mere comparative inspection of his annual reports would reveal to anyone otherwise uninformed how much the Institute owes to him. During his term of office the membership of the Institute has approximately doubled and the financial position of the Institute has been markedly improved in spite of war conditions. For both of these favorable circumstances he deserves the greater part of the credit. In his present position as Chief Statistical Consultant with the Office of Production Research and Development of the War Production Board in association with Holbrook Working, he is still serving the cause of statistics as well as the war effort. The Institute will greatly benefit by the intensive work these men are doing in educating industry in the uses of statistical methods in the control of quality in production.

The annual election of the Institute just concluded by mail resulted in the election of the following officers for 1944: W. A. Shewhart, President; W. G. Cochran and Will Feller, Vice-Presidents; and P. S. Dwyer, Secretary-Treasurer. Professor Dwyer had been appointed Acting Secretary-Treasurer upon the resignation of Professor Olds. I need not remind the members of the Institute that the year 1944 remains critical for the Institute and that these new officers will need the fullest support.

C. C. Craig,
President, 1943.

February 15, 1944.

ANNUAL REPORT OF THE SECRETARY-TREASURER OF THE INSTITUTE

During 1943 three meetings of the Institute were held. On May 29, the Institute met jointly with the American Society of Mechanical Engineers at the Engineering Societies Building in New York City, the program having been arranged by Abraham Wald and A. I. Peterson. A meeting consisting of three evening sessions was held on June 17–19, at George Washington University, Washington, D. C. William G. Madow made the necessary preparations for this meeting. On September 12–13, the Institute held its sixth summer meeting at New Jersey College for Women, Rutgers University, New Brunswick, New Jersey. This meeting was held in conjunction with the summer meetings of the American Mathematical Society and the Mathematical Association of America, and the program was arranged by Abraham Wald. All three of these meetings were well attended.

Early in the year a petition was granted for the establishment of a Pittsburgh Chapter of the Institute. This organization, formerly known as the Society of Quality Control Statisticians, held its first meeting under the auspices of the Institute on June 19, and a second meeting was held on October 9.

As recorded in an addendum to last year's report, Vice-President E. L. Dodd died on January 9, 1943, and Dr. W. E. Deming was appointed to fill out the unexpired term.

Professor H. L. Rietz died on December 7, 1943. A statement of appreciation for his work in mathematical statistics and in connection with the Institute has been prepared by Professor A. R. Crathorne and appears in the present issue of the *Annals*.

The following financial statement covers a period from December 10, 1942, to December 21, 1943 (the books and records of the Treasurer have been audited by Paul S. Dwyer and found to be in agreement with the statement as submitted):

FINANCIAL STATEMENT

December 10, 1942, to December 21, 1943

Receipts

Balance on Hand, December 10, 1942	$2,155.13
Dues	2,640.62
Subscriptions	1,631.67
Sales of Back Numbers	901.47
Miscellaneous	6.07
Total Receipts	$7,334.96

Expenditures

Annals Office		$5.25
Waverly Press		
Printing and Mailing *Annals*—4 issues		2,763.28
Back Numbers Office		
Purchase of back numbers from H. C. Carver	$171.56	
Reprinting 200 copies of Vol. VII, No. 1	108.08	
		279.64
Library Committee		36.11
Secretary-Treasurer's Office		
Printing and Supplies	$135.00	
Binding	2.25	
Postage	135.77	
Clerical Help	171.20	
		444.22
Programs for Meetings		44.22
Board of Directors		41.74
Miscellaneous		5.45
Total Expenditures		$3,619.91
Balance on Hand, December 21, 1943		3,715.05
		$7,334.96

In comparison with the financial condition of the Institute at the end of 1942, the receipts from dues and subscriptions have increased nearly $700 and the proceeds from sales of back numbers have decreased nearly $500. In spite of increased prices, it was possible to reduce expenditures by approximately $800. Thus the Institute finds itself in a somewhat more favorable position than at the end of last year.

Edwin C. Olds,
Secretary-Treasurer.

December 27, 1943.

CONSTITUTION
OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

ARTICLE I

Name and Purpose

1. This organization shall be known as the Institute of Mathematical Statistics.
2. Its object shall be to promote the interests of mathematical statistics.

ARTICLE II

Membership

1. The membership of the Institute shall consist of Members, Junior Members, Fellows, Honorary Members, and Sustaining Members.

2. Voting members of the Institute shall be (a) the Fellows, and (b) all others, Junior Members excepted, who have been members for twenty-three months prior to the date of voting.

3. No person shall be a Junior Member of the Institute for more than a limited term as determined by the Committee on Membership and approved by the Board of Directors.

ARTICLE III

Officers, Board of Directors, and Committee on Membership

1. The Officers of the Institute shall be a President, two Vice-Presidents, and a Secretary-Treasurer. The terms of office of the President and Vice-Presidents shall be one year and that of the Secretary-Treasurer three years. Elections shall be by majority ballots at Annual Meetings of the Institute. Voting may be in person or by mail.

(a) Exception. The first group of Officers shall be elected by a majority vote of the individuals present at the organization meeting, and shall serve until December 31, 1936.

2. The Board of Directors of the Institute shall consist of the Officers, the two previous Presidents, and the Editor of the Official Journal of the Institute.

3. The Institute shall have a Committee on Membership composed of three Fellows. At their first meeting subsequent to the adoption of this Constitution, the Board of Directors shall elect three members as Fellows to serve as the Committee on Membership, one member of the Committee for a term of one year, another for a term of two years, and another for a term of three years. Thereafter the Board of Directors shall elect from among the Fellows one member annually at their first meeting after their election for a term of three years. The president shall designate one of the Vice-Presidents as Chairman of this Committee.

ARTICLE IV

Meetings

1. A meeting for the presentation and discussion of papers, for the election of Officers, and for the transaction of other business of the Institute shall be held annually at such time as the Board of Directors may designate. Additional meetings may be called from time to time by the Board of Directors and shall be called at any time by the President upon written request from ten Fellows. Notice of the time and place of meeting shall be given to the membership by the Secretary-Treasurer at least thirty days prior to the date set for the meeting. All meetings except executive sessions shall be open to the public. Only papers accepted by a Program Committee appointed by the President may be presented to the Institute.

2. The Board of Directors shall hold a meeting immediately after their election and again immediately before the expiration of their term. Other meetings of the Board may be held from time to time at the call of the President or any two members of the Board. Notice of each meeting of the Board, other than the two regular meetings, together with a statement of the business to be brought before the meeting, must be given to the members of the Board by the Secretary-Treasurer at least five days prior to the date set therefor. Should other business be passed upon, any member of the Board shall have the right to reopen the question at the next meeting.

3. The Committee on Membership shall hold a meeting immediately after the annual meeting of the Institute. Further meetings of the Committee may be held from time to time at the call of the Chairman or any member of the Committee provided notice of such call and the purpose of the meeting is given to the members of the Committee by the Secretary-Treasurer at least five days before the date set therefor. Should other business be passed upon, any member of the Committee shall have the right to reopen the question at the next meeting.

4. At a regularly convened meeting of the Board of Directors, four members shall constitute a quorum. At a regularly convened meeting of the Committee on Membership, two members shall constitute a quorum.

ARTICLE V

Publications

1. The *Annals of Mathematical Statistics* shall be the Official Journal for the Institute. The Editor of the *Annals of Mathematical Statistics* shall be a Fellow appointed by the Board of Directors of the Institute. The term of office of the Editor may be terminated at the discretion of the Board of Directors.

2. Other publications may be originated by the Board of Directors as occasion arises.

ARTICLE VI

Expulsion or Suspension

1. Except for non-payment of dues, no one shall be expelled or suspended except by action of the Board of Directors with not more than one negative vote.

ARTICLE VII

Amendments

1. This constitution may be amended by an affirmative two-thirds vote at any regularly convened meeting of the Institute provided notice of such proposed amendment shall have been sent to each voting member by the Secretary-Treasurer at least thirty days before the date of the meeting at which the proposal is to be acted upon. Voting may be in person or by mail.

BY-LAWS

ARTICLE I

Duties of the Officers, the Editor, Board of Directors, and Committee on Membership

1. The President, or in his absence, one of the Vice-Presidents, or in the absence of the President and both Vice-Presidents, a Fellow selected by vote of the Fellows present, shall preside at the meetings of the Institute and of the Board of Directors. At meetings of the Institute, the presiding officer shall vote only in the case of a tie, but at meetings of the Board of Directors he may vote in all cases. At least three months before the date of the annual meeting, the President shall appoint a Nominating Committee of three members. It shall be the duty of the Nominating Committee to make nominations for Officers to be elected at the annual meeting and the Secretary-Treasurer shall notify all voting members at least thirty days before the annual meeting. Additional nominations may be sub-

mitted in writing, if signed by at least ten Fellows of the Institute, up to the time of the meeting.

2. The Secretary-Treasurer shall keep a full and accurate record of the proceedings at the meetings of the Institute and of the Board of Directors, send out calls for said meetings and, with the approval of the President and the Board, carry on the correspondence of the Institute. Subject to the direction of the Board, he shall have charge of the archives and other tangible and intangible property of the Institute, and once a year he shall publish in the *Annals of Mathematical Statistics* a classified list of all Members and Fellows of the Institute. He shall send out calls for annual dues and acknowledge receipt of same; pay all bills approved by the President for expenditures authorized by the Board or the Institute; keep a detailed account of all receipts and expenditures, prepare a financial statement at the end of each year and present an abstract of the same at the annual meeting of the Institute after it has been audited by a Member or Fellow of the Institute appointed by the President as Auditor. The Auditor shall report to the President.

3. Subject to the direction of the Board, the Editor shall be charged with the responsibility for all editorial matters concerning the editing of the *Annals of Mathematical Statistics*. He shall, with the advice and consent of the Board, appoint an Editorial Committee of not less than twelve members to co-operate with him; four for a period of five years, four for a period of three years, and the remaining members for a period of two years, appointments to be made annually as needed. All appointments to the Editorial Committee shall terminate with the appointment of a new Editor. The Editor shall serve as editorial adviser in the publication of all scientific monographs and pamphlets authorized by the Board.

4. The Board of Directors shall have charge of the funds and of the affairs of the Institute, with the exception of those affairs specifically assigned to the President or to the Committee on Membership. The Board shall have authority to fill all vacancies ad interim, occurring among the Officers, Board of Directors, or in any of the Committees. The Board may appoint such other committees as may be required from time to time to carry on the affairs of the Institute.

5. The Committee on Membership shall prepare and make available through the Secretary-Treasurer an announcement indicating the qualifications requisite for the different grades of membership.

ARTICLE II

Dues

Members shall pay five dollars at the time of admission to membership and shall receive the full current volume of the Official Journal. Thereafter, Members shall pay five dollars annual dues. The annual dues of Junior Members shall be two dollars and fifty cents. The annual dues of Fellows shall be five dollars. The annual dues of Sustaining Members shall be fifty dollars. Honorary Members shall be exempt from all dues.

(a) Exception. In the case that two Members of the Institute are husband and wife and they elect to receive between them only one copy of the Official Journal, the annual dues of each shall be three dollars and seventy-five cents.

2. Annual dues shall be payable on the first day of January of each year.

3. The annual dues of a Fellow, Member, or Junior Member include a subscription to the Official Journal. The annual dues of a Sustaining Member include two subscriptions to the Official Journal.

4. It shall be the duty of the Secretary-Treasurer to notify by mail anyone whose dues

may be six months in arrears, and to accompany such notice by a copy of this Article. If such person fail to pay such dues within three months from the date of mailing such notice, the Secretary-Treasurer shall report the delinquent one to the Board of Directors, by whom the person's name may be stricken from the rolls and all privileges of membership withdrawn. Such person may, however, be re-instated by the Board of Directors upon payment of the arrears of dues.

ARTICLE III

Salaries

1. The Institute shall not pay a salary to any Officer, Director, or member of any committee.

ARTICLE IV

Amendments

1. These By-Laws may be amended in the same manner as the Constitution or by a majority vote at any regularly convened meeting of the Institute, if the proposed amendment has been previously approved by the Board of Directors.

If
tice,
hom
ith-
pay-

om-

y a
nd-